Improving Network Fairness With Edge Addition

Mihir Patel

May 15, 2025

Version: Final Draft

Advisor: Dr. Sorelle Friedler

Abstract

In social networks, information, opportunities, and resources spread along the edges of the graph. As a result, some individuals (nodes) may be more advantaged than others purely due to their position in the network. In this work, we formulate several NP-hard optimization problems that aim to quantify and reduce these advantage gaps. Our main contribution is the introduction of the Closeness Ratio Improvement problem, which asks: given a graph and two specified nodes, how can we add a bounded number of edges to make the closeness centralities of the two nodes as similar as possible (specifically, bring their ratio close to 1)? We show that Closeness Ratio Improvement is NP-hard for any target ratio in the interval $(\frac{1}{2}, 1]$. We also provide a trivial $\frac{1}{2}$-approximation algorithm for this problem and demonstrate that several intuitive approximation strategies can perform arbitrarily poorly compared to the optimal solution. Our approach integrates perspectives from both the network editing and network fairness literatures. Finally, we introduce two new problems that extend our fairness framework to focus on equality between groups of nodes and across all node pairs in a graph.

Contents

1 Introduction **2**

1.1 Network Fairness 2

1.2 Group Fairness 3

1.3 Motivation 4

2 Literature Review **6**

2.1 Diameter Minimization 6

2.1.1 NP-Hardness 7

2.1.2 Gonzalez's 2-Approximation for k-CENTER 10

2.1.3 Li's 4-Approximation for DIAMETER MINIMIZATION 12

2.2 Average Shortest Path Length Minimization 15

2.3 Network Fairness 18

3 Statement of The Problem **22**

3.1 Problem Motivation 22

3.1.1 Closeness Centrality 22

3.1.2 Ratio 24

3.2 Problem Formulation 25

4 Hardness of Closeness Ratio Improvement **26**

5 $\frac{1}{2}$-Approximation for CLOSENESS RATIO IMPROVEMENT **30**

6 Intuition-Building Examples **33**

7 Conclusion **36**

A Deferred Proofs From Chapter 4 **38**

A.0.1 Proof of Claim 4.0.1 38

A.0.2 Proof of Claim 4.0.2 39

A.0.3 Remainder of the Proof of Theorem 4.0.2 40

1

Introduction

1.1 Network Fairness

A graph is a mathematical object defined by a set of vertices/nodes V and edges/links $E \subseteq V \times V$ (e.g. Figure 1.1). 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* (boyd et al. 2014)— if LinkedIn users share resources

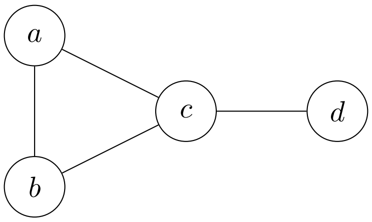

Fig. 1.1.: An example of a graph, $V = \{a, b, c, d\}, E = \{(ab), (ac), (bc), (cd)\}$

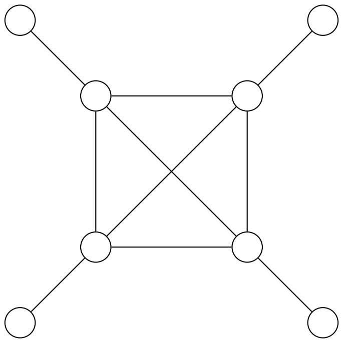

Fig. 1.2.: In this social network, the four central nodes are more advantaged than the four peripheral nodes as they have greater access to the spread of information across the network.

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 1.2, 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 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 (which will be introduced in Chapter 3).

1.2 Group Fairness

With the goal of accurately modeling real-world social networks, a logical area of interest for the network fairness community is *group* network fairness. Returning to the LinkedIn example, users all possess group identities (e.g. race, gender, age group, income bracket) and analyzing all users as identical entities may miss larger group behaviors at play. Within a graph context, we typically view groups as a partition of the set of vertices V. That is, each vertex will belong to only one group, and no vertex can belong to no groups. We can then analyze the previous

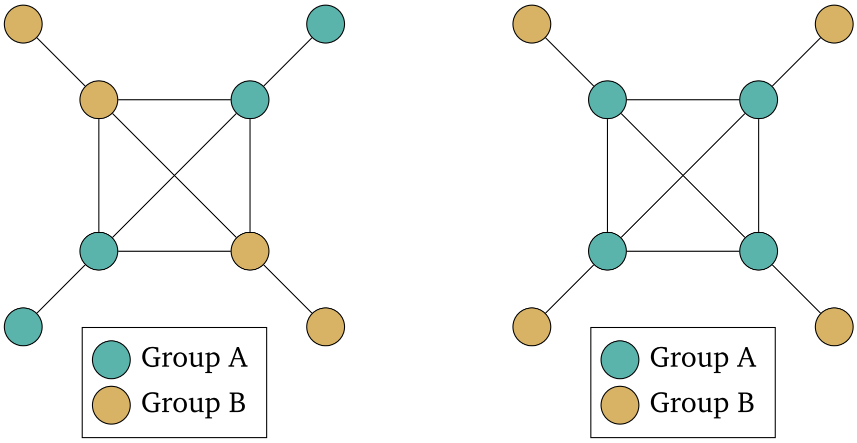

Fig. 1.3.: The network on the left has an advantage gap between nodes, but not an advantage gap between groups. The network on the right has an advantage gap between nodes and and advantage gap between groups.

network (un)fairness example, but this time in a group setting. Still, individual nodes will have more advantage than others in the graph. However, in Figure 1.3, one partitioning of the vertices does not have an advantage gap at a group level, while the other does. In the network on the left, of the four advantaged vertices, two belong to group A and two belong to group B. Similarly, of the four disadvantaged vertices, two belong to each group. At each level of advantage, vertices are split equally between groups. On the other hand, in the network on the right, all four advantaged vertices are in group A, while all four disadvantaged vertices are in group B. In this case, group A as a whole is more advantaged than group B, and this now presents a group advantage gap in addition to the individual node advantage gap.

1.3 Motivation

The proposed work seeks to reduce advantage gaps between nodes and/or groups by adding edges into a graph. This is a relatively new area of research — there exists distinct research in both network-editing algorithms and network fairness, yet the intersection of these fields is less explored (although Bashardoust et al. 2023 and Bhaskara et al. 2024 represent early efforts to address this gap). This work aims to build upon this existing research, from creating definitions of what it means to be fair, to applying and modifying known algorithmic strategies to these newly phrased optimization problems. These problems are non-trivial (NP-hard, in fact), and desired algorithmic results will be mainly theoretical, as provable approximation bounds are of great importance in practice. For example, understanding which connections create fair social structures is extremely useful for equitable follower recommendation algorithms on platforms such as LinkedIn, where greater connectivity often leads to greater access to job opportunities.

In more specific terms, I am interested in the problem of adding edges to a graph in order to equalize the closeness centrality of two vertices. Closeness centrality is a node-level statistic which represents how close a vertex is to all other vertices, or in other words, how "important" a vertex is in the graph. At a high level, our problem, CLOSENESS RATIO IMPROVEMENT, seeks to add a budgeted number of edges to a graph, with the goal of making two vertices equally important in the graph (for a more formal problem statement, see Chapter 3).

2

Literature Review

With the goal of ultimately developing algorithms and proving hardness of the problem proposed above, I take an in-depth look at algorithms which add edges to a graph to reduce its diameter (DIAMETER MINIMIZATION). Compiling results from several papers, I present proofs of NP-hardness and a 4-approximation for this problem. Then, I look at the closely related problem of adding edges to a graph in order to minimize the average shortest path distance, and briefly present various problem statements and an overview of algorithmic results. Finally, I look at two papers which propose different definitions of node advantage in a graph in the independent cascade model. By combining existing theoretical results with these better, yet mathematically more complex fairness definitions, I hope to inform further research in the proposed area.

2.1 Diameter Minimization

The problem of minimizing the diameter of a graph is well studied. Specifically, the BOUNDED CARDINALITY MINIMAL DIAMETER variant of this problem, which we will refer to as DIAMETER MINIMIZATION, seeks to add a specified number of edges to the graph in order to reduce the diameter as much as possible. Formally, we define the problem as follows. Note that the shortest path length between two vertices $u, v \in V$ within a graph $G = (V, E)$ is represented as $\mathrm{d}_G(u, v)$, and a graph $G = (V, E)$ augmented with an edge set $S \subseteq V^2 - E$ is denoted as $G + S = (V, E \cup S)$.

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 $S \subseteq V^2 - E$ of size at most k such that $D_{G+S} = \max_{u,v \in V} \mathrm{d}_{G+S}(u, v)$ is minimized.

Why is this relevant for our problem? 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 Diameter Minimization perform in improving the closeness centrality of a vertex.

Although many such strategies exist, we focus on approximation algorithms which use the k-Center problem (formalized in Section 2.1.2). 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. Originally, this was thought to be a 4-approximation (Li et al. 1991), but tighter analysis has shown this strategy in fact achieves a 2-approximation (Bilò et al. 2012). In the ensuing sections, we present a proof for the NP-hardness of Diameter Minimization, then show Gonzalez's 2-approximation for k-Center and its use in Li's 4-approximation for Diameter Minimization.

2.1.1 NP-Hardness

We now analyze an NP-hardness proof of Diameter Minimization, which reduces from Set Cover (Adriaens and Gionis 2022). Although many hardness proofs for Diameter Minimization exist, this one reduces from a familiar problem whose hardness is well-established (Karp 1972) and does not require additional justification.

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, we also formulate DIAMETER MINIMIZATION as a decision problem.

DIAMETER MINIMIZATION (DECISION VARIANT)

Input: A graph $G = (V, E)$ and a positive integer $k \in \mathbb{N}$.

Problem: Is there a set of edges $S \subseteq V^2 - E$ of size at most k such that $D_{G+S} \leq \delta$?

Theorem 2.1.1 (Adriaens and Gionis 2022). DIAMETER MINIMIZATION *is* NP*-Hard.*

Proof Idea: We reduce from SET COVER, constructing a graph (Figure 2.1) where all vertices in a $2k+1$-sized clique are 4 away from all vertices which represent elements from our SET COVER instance. We show that if there is a set cover of size k, we can create k shortcut edges to set vertices of the cover, achieving diameter 3. Inversely, if there is no set cover of size k, there must remain one element vertex which is 4 away from at least one vertex in the $2k+1$-sized clique.

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 $k \in \mathbb{N}$, we construct a decision instance of DIAMETER MINIMIZATION. Create a clique of vertices $\{x_1, x_2, \ldots, x_{2k+1}\}$, and connect all of them to the vertex a, we will 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 $e_1, e_2, \ldots, e_n$ 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). We now have an instance of DIAMETER MINIMIZATION, $(G = (V, E), k)$. The construction of G is depicted in Figure 2.1. 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 $D_{G+T} \leq 3$.

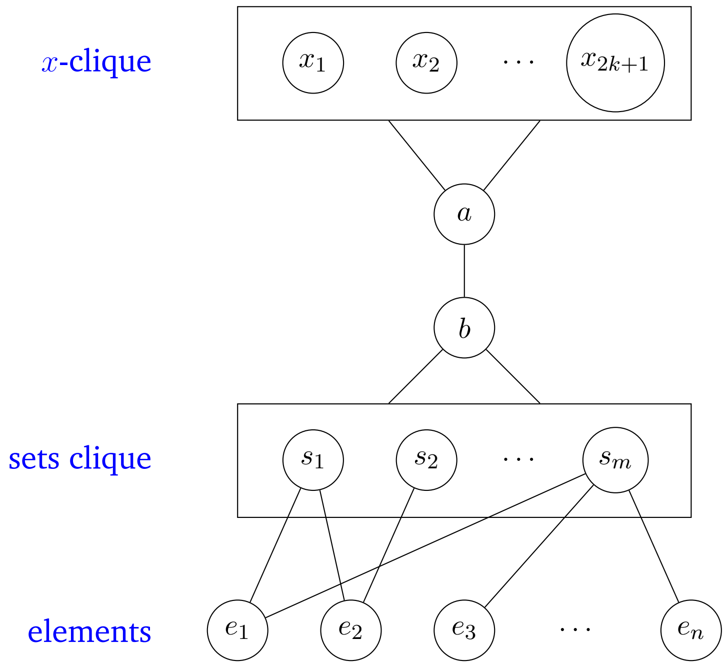

Fig. 2.1.: Construction of a Diameter Minimization instance from Set Cover (vertices within a box form a clique).

Before any edge additions, the diameter of G is 4. The only paths of length 4 are paths between vertices in X and element vertices. Thus to decide if $D_{G+T} \leq 3$, we only need to consider if all of these paths are reduced to length 3.

k-sized set cover $\Longrightarrow$ 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 $a - s_i$. 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 $\Longrightarrow$ 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 any set vertex adjacent to it. We can show this by contradiction — suppose every element vertex *did* have a new edge incident on it or one of the set vertices containing it. For any element vertex that is the endpoint of a new edge, simply reconstruct this edge to go to one of the sets containing this element instead. Now we have identified k edges incident only on set vertices such that every element is the neighbor of at least one such set vertex, and thus we have identified 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. We also know that there must 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). Let's trace the path from e to x. e must go to the set level, at which point it arrives at some vertex s containing it (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, we know 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 $\mathrm{d}_{G+T}(e, x) > 3$ and thus $D_{G+T} > 3$, as desired.

□

With a simple reduction from SET COVER, we are able to show that the *decision variant* of DIAMETER MINIMIZATION is NP-hard. This further shows that the optimization variant (the problem we hope to eventually solve) is NP-hard as well. Take an instance of DIAMETER MINIMIZATION with a δ, for which we must decide if we can get the diameter less than δ with k edge additions. If you can solve the optimization variant, solve it on this problem instance. If the diameter you get is less than δ, return YES, otherwise return NO. You have now solved the decision variant of DIAMETER MINIMIZATION (which we know you cannot do), implying the optimization variant is also NP-hard.

Showing that DIAMETER MINIMIZATION is NP-hard re-positions research goals for the problem, as trying to design an optimal algorithm is now at least equivalent to trying to prove P=NP (if not harder). We thus present approximation results which rely on Gonzalez's approximation for k-CENTER, presented below.

2.1.2 Gonzalez's 2-Approximation for k-CENTER

The k-CENTER problem (specified for graphs) seeks to identify k center vertices which minimize the maximum distance any vertex is from its closest center. After defining the problem formally, we present a simple greedy 2-approximation for this problem. 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 $\max_{v \in V} \min_{c \in C} \mathrm{d}_G(v, c)$ is minimized.

R_C denotes 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 $|C| \leq k$, $R_C = \max_{v \in V} \min_{c \in C} \mathrm{d}_G(v, c)$. We use C^* to denote the optimal set of centers. Specifically, $C^* = \arg\min_{C \subseteq V, |C| \leq k} R_C$.

Algorithm (Gonzalez 1985)

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. 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, we need to 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 we need to do this for k iterations, meaning this approach takes $O(k^2 n)$ time overall. We can optimize this algorithm by maintaining an array of n elements that stores each vertex's distance to its closest center and in each iteration, updating an entry if the new center we added reduces this distance. Then each iteration takes $O(n)$ (find the vertex v with the largest value in the array, calculate the distance of each vertex to v, update at most n elements if necessary). Overall, this optimized version would take $O(kn)$, assuming a distance oracle, but we omit it for sake of simplicity.

Theorem 2.1.2 (Gonzalez 1985). *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. Suppose we have a graph $G = (V, E)$ and $k \in \mathbb{N}$. In order to show existence, we propose 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

Algorithm 1 Gonzalez's 2-Approximation for k-CENTER

```
procedure GONZALEZ($G = (V, E), k$)
    Choose $x \in V$ arbitrarily
    Let $C = \{x\}$
    while $|C| < k$ do
        Let $w$ be $\arg\max_{v \in V} \min_{c \in C} \mathrm{d}_G(v, c)$, breaking ties arbitrarily
        $C \leftarrow C \cup \{w\}$
    end while
    return $C$
end procedure
```

from the construction of our algorithm — C is exclusively selected from V in an iterative fashion until $|C| = k$.

We now prove 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). We will 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 $\mathrm{d}_G(u, v) \geq R_C$. 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 tells us that either $\mathrm{d}_G(u, c) \geq \frac{R_C}{2}$ or $\mathrm{d}_G(v, c) \geq \frac{R_C}{2}$. If neither were true, there would be a path between u and v of length less than R_C, which we have established is not possible. In any case, we have shown that 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. □

To find an approximate solution for k-CENTER, *Gonzalez* is a very simple greedy algorithm — iteratively choose the vertex which maximizes our center-distance, adding it to our set of centers. This approach provably achieves a 2-approximation.

2.1.3 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, $S^* = \arg\min_{S \subseteq V^2 - E, |S| \leq k} D_{G+S}$.

Algorithm (Li et al. 1991)

Given an instance of DIAMETER MINIMIZATION with graph G and k edges, 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 we remain within our budget of k for DIAMETER MINIMIZATION, even though we used $Gonzalez$ as a subroutine with $k + 1$ centers.

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.

Algorithm 2 Li's 4-Approximation for DIAMETER MINIMIZATION

```
procedure LI($G = (V, E), k$)
    Let $C$ be the set of centers outputted by $Gonzalez(G, k + 1)$
    Fix $x$, an arbitrary vertex in $C$
    Let $S$ be the set of edges not already in $G$ which connect $x$ to each other center in $C$
    return $S$
end procedure
```

We now provide an example run of Li's algorithm for DIAMETER MINIMIZATION on Figure 2.2 with $k = 2$. We will refer to Figure 2.2 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 will show, 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.

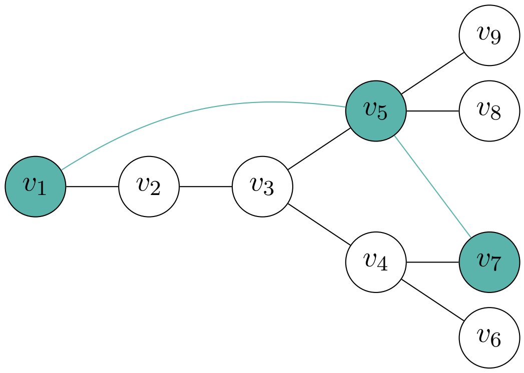

Fig. 2.2.: Example run of Li's 4-approximation for DIAMETER MINIMIZATION. Blue vertices indicate outputted centers from *Gonzalez*, blue edges are proposed edge additions from *Li*.

Lemma 2.1.1 (Li et al. 1991)**.** *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^*, we will show how to construct a solution C 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^*, we added at most one center to C for each edge, and we instantiated C with one vertex, thus we have $|C| \leq k+1$.

Now we want to show 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 $G + S^*$, which would imply that $R_C \leq \mathrm{d}_{G+S^*}(v, x) \leq D_{G+S^*}$. Consider the shortest path from v to x in $G + S^*$. We have two cases: either this path uses an edge in S^* or it does not. In the first case, we will call the first new edge it encounters $(ab) \in S^*$ and, without loss of generality, $\mathrm{d}_G(v, a) \leq \mathrm{d}_G(v, b)$. Then by construction of C, we know that 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, we know 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 2.1.3 (Li et al. 1991). *Given $G = (V, E)$ and $k \in \mathbb{N}$, there exists a polynomial-time algorithm which produces $S \subseteq V^2 - E$ such that $|S| \leq k$ and $D_{G+S} \leq 4D_{G+S^*} + 2$.*

Proof. Suppose we have a graph $G = (V, E)$ and $k \in \mathbb{N}$. In order to show existence, we propose the polynomial-time algorithm Li, which outputs a set of edges $S \subseteq V^2 - E$ such that $|S| \leq k$ and $D_{G+S} \leq 4D_{G+S^*} + 2$. The first two claims follow immediately from the construction of our 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 $\mathrm{d}_{G+S}(u, v) \leq 2R_C + 2$ for any $u, v \in V$, implying $D_{G+S} \leq 2R_C + 2$. We know that $Gonzalez$ can be used to achieve a 2-approximation for the k-Center problem, so $D_{G+S} \leq 2(2R_{C^*}) + 2$, and then by the above lemma, $D_{G+S} \leq 2(2D_{G+S^*}) + 2 = 4D_{G+S^*} + 2$, as desired. □

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.

2.2 Average Shortest Path Length Minimization

A variant of Diameter Minimization is the problem of minimizing the average shortest path length in a graph, formulated in (Meyerson and Tagiku 2009). This problem is much less explored than Diameter Minimization, so there are much fewer algorithmic strategies/results/improvements to draw from. However, as we are concerned with making a small number of vertices closer to all vertices, this problem seems more relevant than Diameter Minimization, which minimizes the *maximum* vertex-pair distance in a graph. As a result, we provide a results summary of this paper, as opposed to the more in-depth proofs from the previous section.

Meyerson and Tagiku introduce several different problem statements for minimizing the average shortest path distance, and draw analogs between this problem and the k-Median with Penalties problem. We focus on two subproblems, minimizing the single-source average shortest path distance and minimizing the unweighted all-pairs average shortest path distance (both presented below). When considering how to best improve the closeness centrality of a select number of vertices, using strategies which focus on making one vertex closer to all other vertices are intuitively more applicable than strategies which are concerned with compressing the entire graph (the initial reason we looked at this paper in addition to conventional diameter minimization literature). Similarly, as our problem is in an unweighted graph setting, looking at the unweighted formulation of average shortest path distance minimization will provide clearer, transferable algorithmic intuiton. Ultimately, Meyerson and Tagiku show that if α is the best approximation factor for k-Median with Penalties, they can obtain an α-approximation for minimizing the average shortest path distance from a single vertex, and a 2α-approximation for minimizing the all-pairs unweighted analog.

First, we formalize the k-Median with Penalties problem, which will be used as a subroutine in their algorithms. Given a set of cities C and potential facility locations F in a metric space, each city $c \in C$ has an associated demand w_c and penalty cost p_c. We can choose to serve a city or not, and we have binary variable x_c as 1 if we are serving that city, 0 if not. We want to find the k facilities to open such that our overall cost (defined as the sum over all cities of distance to its closest facility times demand, or that city's penalty cost, depending on which is smaller) is minimized.

Metric k-Median With Penalties

Input: A metric space $(C \cup F, d)$, $k \in \mathbb{N}$ and a demand w_c and penalty p_c for all $c \in C$..

Problem: Find a set of facilities $F' \subseteq F$ such that $|F'| \leq k$ and $\sum_{c \in C}((\min_{f \in F'} d_{cf} w_c)x_c + p_c(1 - x_c))$ is minimized.

Now, we provide the formulations of our relevant problems. Note that the pairwise distances between elements i, j is now denoted as d_{ij}, instead of $\mathrm{d}_G(i, j)$, which is how we denoted vertex distances for Diameter Minimization in the graph space.

This notation represents our containment within an arbitrary metric space, and is consistent with the use of subscripts for other variables in this problem.

For our problem instance, let $G = (V, E)$ be an undirected graph with non-negative edge lengths l_e for each $e \in E$ (conventionally defined of as edge weights) and non-negative vertex pair weights w_{uv} for each $u, v \in V$ (conventionally defined as vertex-pair costs). Define d_{uv} as the length of the shortest path between $u, v \in V$. For a vertex u, the *weighted one-to-all shortest path sum* in G is defined as $D_u(G) = \sum_{v \in V} w_{uv} d_{uv}$, the sum over all vertices of the distance from u times the weight between u. The *weighted all-pairs shortest path sum* is $D(G) = \sum_{u \in V} D_u(G)$, the sum of one-to-all shortest path sums over all vertices. The *unweighted all-pairs shortest path sum* is defined over the same vertices, just excluding weights (i.e. $\sum_{u \in V} \sum_{v \in V} d_{uv}$).

Single Source Average Shortest Path Distance Minimization (SS-ASPDM)

Input: A graph $G = (V, E)$, a vertex $u \in V$ and positive integers $k, \delta \in \mathbb{N}$.

Problem: Find a set of edges $S \subseteq V^2 - E$ which minimizes $D_u(G + S) = \sum_{v \in V} (w_{uv} d_{uv})$, such that $|S| \leq k$ and $l_s \leq \delta$ for all $s \in S$.

Unweighted Average Shortest Path Distance Minimization (U-ASPDM)

Input: A graph $G = (V, E)$, a vertex $u \in V$ and positive integers $k, \delta \in \mathbb{N}$.

Problem: Find a set of edges $S \subseteq V^2 - E$ which minimizes $D(G + S) = \sum_{u \in V} \sum_{v \in V} d_{uv}$, such that $|S| \leq k$ and $l_s \leq \delta$ for all $s \in S$.

Note that both of these problem give an number-of-edges budget (k), as well as a length-of-edges budget (δ), which is different than Diameter Minimization (where all edges had length 1, so δ was unnecessary).

For any instance of SS-ASPDM (say we are trying to minimize the average shortest path distance of vertex u), their key observation is that an optimal solution will only include edges incident on u. This is fairly intuitive, for any other edge ab, the edge ub will always do a better job of getting u close to b (and every other vertex whose path from u uses b). With this claim in mind, the task of SS-ASPDM is to then identify the k vertices to connect u to such that we minimize our cost function, $D_u(G) = \sum_{v \in V} w_{uv} d_{uv}$. They then show that this can be re-formulated as an instance of k-Median with Penalties, i.e. the k vertices to connect to are the k

facilities to open, so that we are minimizing the distance from cities (non-facility vertices) to their closest facility. Thus SS-ASPDM can be approximated with the same factor as k-Median with Penalties.

For U-ASPDM, they argue that there must be some vertex u for which running the SS-ASPDM algorithm (i.e. formulating as k-Median with Penalties) does boundedly well, specifically a 2α-approximation. Specifically, for the optimal set of edges S^*, there must be some vertex u with $D_u(G + S^*) \leq \frac{1}{n}D(G + S)^*$ (simply because $D(G + S^*)$ is the sum of SS-ASPDM values over n vertices). Then we know that we can find a set of edges S such that $D_u(G+S) \leq \alpha D_u(G+S^*)$. This claim is worked out in more detail in the paper, but the final main observation is that if $d_{ij} \leq d_{iu}+d_{uj}$ (triangle inequality, as we are in a metric space) and $D(G + S^*) = \sum_{i,j \in V} d_{ij}$, then we can bound $D(G + S) \leq 2nD_u(G + S)$. Putting this all together,

$$D(G+S) \leq 2nD_u(G+S) \leq 2n\alpha D_u(G+S^*) \leq 2n\alpha \frac{1}{n} D(G+S^*) = 2\alpha D(G+S^*)$$

This existence proof requires construction from a vertex u, derived from knowledge of the optimal set of edges S^*. In implementation, without access to the optimal set of edges, finding u would require guessing and trying all vertices, and choosing the one for which the algorithm for SS-ASPDM returns the smallest average shortest path distance, but this will still be polynomial time.

The algorithmic approach, and associated approximation bound for U-ASPDM is a particularly interesting result — we can make all edge additions incident onto one vertex and still get a constant-factor approximation for minimizing the average shortest path distance over *all* vertices. While average shortest path minimization is less well-researched, it higher level goal is arguably more similar to closeness centrality ratio improvement.

2.3 Network Fairness

We argue that social networks can model the spread of information across a network (as motivated in the Introduction). If this is the case, operating under the assumption that information is guaranteed to flow between two individuals if they are connected seems an overstep. In other words, using shortest-path definitions to quantify how likely information is to spread between two nodes does not fully capture how information flows. For example, if two nodes a, b are connected by one path of length 100, we may say that they are far apart, and thus it is unlikely information may flow between them (in the shortest-path metric). However, if they are connected with 1,000 paths of length 101, while these paths are not the *shortest* path connecting a, b, they still can carry information (and there are a lot of them). Overall, a and b feel highly connected, and our shortest-path definitions of information access (diameter, eccentricity, closeness centrality) fail to capture this.

As a result, we look at Bashardoust et al. 2023 and Fish et al. 2019, which propose numerous definitions to quantify network advantage under the independent cascade (IC) model of information flow. In IC, some set of seed nodes start with "information" and information will be transmitted across edges with a fixed probability α. The access distance p_{ij} is defined as the probability that, if information starts at vertex i, it will reach vertex j. Note that as they deal with undirected graphs and all edges have probability α, $p_{ij} = p_{ji}$. As opposed to typical graph measures, we are not limited to shortest paths. That is, in denser graphs, where there are multiple paths between vertices, the access distance is often greater than just α raised to the length of the shortest path between the vertices (although this is the case in trees, where there is always only one path between vertices). Using the graph in Figure 1.1 as an example, $\alpha^1 \neq p_{ab} = 1 - (1-\alpha)(1-\alpha^2)$, i.e. one minus probability that neither the path $a - b$ and the path $a - c - b$ successfully transmit information from a to b.

Independent Cascade is a model that addresses the issues we put forth in the previous paragraph, as it better captures the spread of information across a network. With this basic definition of access distance in mind, Bashardoust et al. 2023 creates analogs of conventional graph definitions, now in IC. We will go through some of these definitions that may be useful for our problem, and then look at a problem statement from Bashardoust et al. 2023, as well as some algorithms to solve the problem.

The *access diameter* of a graph is the smallest access distance between two vertices in the graph (the "furthest" apart vertices are the vertices with the smallest probability of information reaching each other). The betweenness centrality of vertex i is defined as $\sum_{i \neq j \neq k \in V} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$ where σ_{jk} is the number of shortest j, k paths and $\sigma_{jk}(i)$ is the number of those paths which go through i. They define the *access centrality* of i as $\sum_{i \neq j \neq k \in V} \frac{p_{jk}(i)}{p_{jk}}$, where $p_{jk}(i)$ is $p_{jk} - p'_{jk}$ with p' being the access probability in the graph with vertex i removed. In simpler terms, access centrality seeks to describe how much of the "information" between all-pairs of vertices is flowing through vertex i.

Next, they define measures of advantage for vertices in a graph. The *broadcast advantage* of a vertex i is its minimum access distance over all other vertices. The *influence advantage* of vertex i is its average access distance over all other vertices. Finally, the *control advantage* of a vertex is simply that vertex's access centrality.

Aside from definitions, this paper introduces the MaxWelfare-Augmentation problem, which is a similar network-editing problem that may be of interest (although this paper does not present any provable algorithmic results).

MAXWELFARE-AUGMENTATION

Input: A graph $G = (V, E)$ and a positive integer $k \in \mathbb{N}$.
Problem: Find a set of edges $S \subseteq V^2 - E$ such that in $G' = (V, E \cup S)$, $\min_{i,j \in V} p_{ij}$ is maximized.

This paper presents some algorithmic results to solve this problem. Although all of these results are empirical, they may still be useful for understanding what kind of strategies can create fairness across a network. Generally, their heursitics are greedy algorithms: find the vertices which realize some minimum, and intervene on them, doing this k times. For example, the heuristic bc-chord, finds and connects the two vertices that realize the minimum access distance across the graph (i.e. the two vertices that are least likely to be able to transmit information), doing this k times. The heuristic infl would find the vertex that has the minimum influence (average access distance to all other vertices) and connect this to the vertex with maximum broadcast (the center). They also tried diam-chord, which finds the vertices which realize the shortest-path diameter, and connects them, doing this k times.

Overall, bc-chord performed the best on a corpus of large social networks in implementation. It improved the minimum broadcast and influence in the graph just as much, if not more, than any other heuristic. In general, broadcast based and influence based strategies performed very similar in this regard. That is, both of them would simultaneously improve the broadcast and influence in the graph, suggesting that these concepts are tied together. Specifically, this implies that nodes that have the worst broadcast in the network also likely have low influence in the network (and vice versa).

Another observation is that bc-chord shrank advantage gaps. While the problem formulation only explicitly considers maximizing the lowest access distance in the graph, if edge augmentations achieve this while simultaneously increasing the gap between the highest and lowest access distance, we do not achieve the equality in the network we hope for. Some of the heuristics would in fact increase this gap, but bc-chord did not. Finally, interventions were more effective in achieving small advantage gaps when the network was better connected initially (whether because α is higher or the graph was naturally more dense). If the former, this bodes well for the success of edge intervention strategies in the shortest-path metric, where edges have associated transmission probabilities of 1 (the highest possible α).

To summarize, Bashardoust et al. 2023 proposes some graph definitions in IC that are of interest for describing overall network fairness. Furthermore, they consider the problem of adding edges to a graph to maximize its minimum pairwise access distance. They find that the strategy of connecting the vertices which realize this minimum performs well empirically in all regards.

(Fish et al. 2019) also deals with graph structures within the IC model. However, most advantage definitions are over subsets of vertices and include a seed set of vertices. While this was technically the case for (Bashardoust et al. 2023), they formulated pairwise access distances where one of the vertices was the only seed – expressing advantage in terms of simpler, pairwise distances that are analagous to conventional graph definitions. Upon achieving algorithmic results similar to those for DIAMETER MINIMIZATION, a next step would to be to derive corresponding results in IC (and the definitions from (Bashardoust et al. 2023) are better positioned for this than (Fish et al. 2019)). Still, there are some interesting behaviors which this paper defines which should be considered when creating problem statements/definitions for what it means to be “fair”. Specifically, it characterizes the “rich get richer” phenomenon, where, in the process of trying to improve the advantage of less-advantaged vertices we inadvertently create a larger gap between the least and most advantaged nodes. This introduces the balance between improving advantage of a vertex and minimizing the advantage gap between vertices, helping decide if proposed algorithms are in fact making social networks more fair in our desired manner.

Most of the algorithmic results from these papers are experimental, so of slightly less interest to our problem. Still, surveying these IC definitions provides useful insights for future directions after algorithmic results may be obtained, and we are able to progress to more difficult problems. Overall, I hope to use algorithmic strategies from Section 2.1 and Section 2.2 to inspire closeness centrality improvement, and if successful, transition to more societally and mathematically complex fairness definitions described in Section 2.3.

3

Statement of The Problem

Equipped with the literature from the previous chapter, we now provide societal and mathematical motivation for certain choices within the problem formulation, concluding with the problem statement itself.

3.1 Problem Motivation

At a high level, we want our problem to require adding edges to a graph to make two nodes equally important in the graph. First, we consider measures of node "importance", and then ways of quantifying how close to equal two nodes are under these measures.

3.1.1 Closeness Centrality

While various measures of centrality (node importance) exist, we have chosen to look at closeness centrality. If the shortest path length between two vertices $u, v \in V$ within a graph $G = (V, E)$ is represented as $\mathrm{d}_G(u, v)$, we define the closeness centrality of a vertex $v \in V$ as $\mathrm{cc}_G(v) = \sum_{u \in V} \mathrm{d}_G(u, v)$, 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 for the most part closer to all other vertices in the graph, i.e. more important. Having a high value implies a vertex is far away from all other vertices in the graph, i.e. less important.

Why have we chosen closeness centrality as our measure of node importance, when many other such measures exist? Intuitively, reducing the closeness centrality of a vertex is similar to reducing the diameter of a graph, or the all-pairs average shortest path length — they all involve adding edges in order to reduce the length of long paths in the graph. Therefore, it is reasonable to assume these well-researched areas of network-editing literature may provide insight into algorithmic strategies for improving the closeness centrality of a vertex. A key similarity between these problems is that adding edges can never be detrimental (you cannot increase the diameter, average shortest path length, or closeness centrality by adding an edge).

This is different than betweenness centrality, the most common (and arguably, the most expressive) measure of centrality. The betweeness centrality of a vertex v is defined as the sum over all non-v vertex pairs of the proportion of shortest paths which go through v, over the number of shortest paths. Formally, it is defined as $b(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} is the number of shortest paths between s and t, and $\sigma_{st}(v)$ is the number of shortest paths between s and t which go through v. When trying to equalize the betweenness centrality of two vertices, adding edges is no longer always beneficial. In fact, adding an edge to increase the betweenness centrality of a vertex can inadvertenly decrease the betweenness centrality of another vertex. For example, if the shortest path between $i, j \in V$ goes through $k \in V$, but an edge addition creates an $i - j$ path shorter than before (and k is not included on this path), then k's betweenness centrality may decrease.

Furthermore, due to a lack of existing results surrounding equalizing the betweenness centrality of two vertices, formulating the problem with closeness centrality (at least initially) is a logical first step. This provides an easier first case that, if significant results are achieved, may give insight into different, more complex centrality measures.

Beyond the above mathematical argument for choosing closeness centrality, a societal argument can also be made. That is, closeness centrality does a better job of capturing certain scenarios of network un-fairness than other centrality measures. Closeness centrality 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 hear this information as soon as possible. Improving a vertex's closeness centrality accomplishes this, whereas improving a vertex's betweenness centrality may leave a small amount of vertices very far from the vertex we hope to improve (thus less likely to receive information if it starts at a random vertex). Betweenness centrality seeks to optimize a proportion and may weigh the importance of minimizing certain path lengths, while closeness centrality seeks to minimize all path lengths from our target vertex. Having now defined (and justified) our measure of node importance, we now define our measure of equality, and then present a formal problem statement.

3.1.2 Ratio

Given vertices a and b, and intuitive way to quantify how close their closeness centralities are to each other would be to take the absolute value of their difference, and try to minimize that value. That is, given a graph G, vertices a and b, and a budget k, find the k edges which (when added to the graph) minimize the absolute value of the closeness centrality of a minus the closeness centrality of b. However, we propose the alternative task of making their ratio of their closeness centrality as close to 1 as possible.

In short, this is because ratio scales better than difference, especially for smaller centrality values. For arithmetic simplicity, we do not normalize our closeness centrality, but most centralities are 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, especially when we analyze centrality measures that follow typical behaviors (Yeh et al. 2024), specifically between 0 and 1, with higher numbers implying a vertex is more central.

3.2 Problem Formulation

Having justified both closeness centrality and ratio measures, we now present our problem. Most of the following definitions have been presented in previous sections, but are included again for clarity. We denote a graph $G = (V, E)$ augmented with an edge set $S \subseteq V^2 \backslash E$ as $G+S = (V, E \cup S)$. Within G, the shortest path length between two vertices $u, v \in V$ is represented as $\mathrm{d}_G(u, v)$. Finally, we define the closeness centrality of a vertex $v \in V$ in a graph $G = (V, E)$ as $\mathrm{cc}_G(v) = \sum_{u \in V} \mathrm{d}_G(u, v)$, 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, and thus more important or central in the graph.

In the problem of Closeness Ratio Improvement, we are given a graph G and vertices a, b, and the goal is to add the k edges to G which will maximize the ratio of a and b's closeness centralities. We define the closeness ratio of a and b as $\text{cc-ratio}_G(a, b) = \frac{\min(\ \mathrm{cc}_{G+S}(a)\ ,\ \mathrm{cc}_{G+S}(b)\ )}{\max(\ \mathrm{cc}_{G+S}(a)\ ,\ \mathrm{cc}_{G+S}(b)\ )}$, so that we can guarantee our ratio is never more than 1.

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 $\text{cc-ratio}_{G+T}(a, b)$.

In the following sections, we first prove that this problem is NP-hard and then present a simple $\frac{1}{2}$-approximation algorithm. We conclude by showing that several common algorithmic strategies fail to surpass the $\frac{1}{2}$-approximation ratio. Unlike the preceding content, which is based on existing research or minor modifications thereof, all results in these sections are original work which is the result of joint research with Dr. Blair Sullivan, Alex Crane, and Aidan Wilde at the University of Utah. These results are further expanded upon in Crane et al. 2025.

4

Hardness of Closeness Ratio Improvement

In this section, we will use a similar constrution as in Adriaens and Gionis 2022 to establish the NP-hardness of Closeness Ratio Improvement by reducing from Set Cover. To do so, we first formally define the decision variant.

Closeness Ratio Improvement (Decision Variant)

Input: A graph $G = (V, E)$, vertices $a, b \in V$, a budget $k \in \mathbb{N}$, and a target ratio $\tau \in (0, 1]$.

Question: Does there exist a set $T \subseteq V^2$ of size at most k such that cc-ratio$_{G+T}(a, b) \geq \tau$?

The following 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.

Theorem 4.0.1. Closeness Ratio Improvement *is* NP*-hard for* $\tau = 1$.

Proof. Let $(U, S_1, S_2, \ldots, S_m, k)$ be an instance of Set Cover, where U denotes a universe of n elements and $\{S_j\}_{j=1}^m$ is a set family over U. We construct an instance $(G = (V, E), a, b, k', \tau = 1)$ of Closeness Ratio Improvement as follows; see Figure 4.1 for a visual depiction. We create three vertices a, b, c, which induce a triangle. We create n *element vertices* $v_1, v_2, \ldots, v_n$ corresponding to the elements $u_1, u_2, \ldots, u_n$ of U, ordered arbitrarily. For each set S_j, we create a *set vertex* s_j, as well as the edge cs_j and the edges $s_j v_i$ for all $u_i \in S_j$. Finally, we create $X = n + k$ vertices, and add an edge from each of these vertices to a. We also set $k' = k$. This completes the construction.

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 cc-ratio$_{G+T}(a, b) = 1$. For the forward direction,

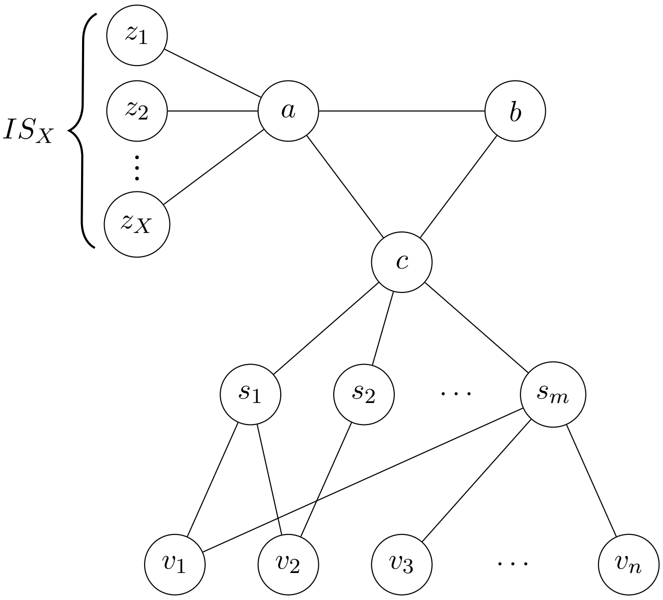

Fig. 4.1.: The construction given by Theorem 4.0.1. We denote set vertices by s_j and element vertices by v_i. Here, IS_X is an independent set of X vertices, with each vertex adjacent to a. We refer to the proof of Theorem 4.0.1 for a formal description of the construction and the accompanying analysis.

suppose that there exists a cover $C \subseteq \{S_j\}_{j=1}^m$ of size k. Let $T = \{bs_j \colon S_j \in C\}$, i.e., the set of edges from b to the set vertices corresponding to sets in the cover C. Then for each element vertex v_i, $\mathrm{d}_{G+T}(v_i, b) = 2$, and $\mathrm{d}_{G+T}(v_i, a) = 3$. Also, b is adjacent to k set vertices, so $\sum_{j=1}^m \mathrm{d}_{G+T}(s_j, b) = 2m - k$, while $\sum_{j=1}^m \mathrm{d}_{G+T}(s_j, a) = 2m$. Then

$$\mathrm{cc}_{G+T}(b) = 2X + 1 + 1 + 2m - k + 2n = 2(n+k) + 2m - k + 2n + 2 = 4n + 2m + k + 2.$$

Similarly, $\mathrm{cc}_{G+T}(a) = X + 1 + 1 + 2m + 3n = 4n + 2m + k + 2$, so $\mathrm{cc}_{G+T}(b) = \mathrm{cc}_{G+T}(a)$, implying cc-ratio$_{G+T}(a, b) = 1 = \tau$, as desired.

For the reverse direction, assume that there is no collection of k sets which covers U. We must prove that there is no set $T \subseteq V^2 \setminus E$ of size at most k such that $\mathrm{cc}_{G+T}(b) = \mathrm{cc}_{G+T}(a)$. We accomplish this by showing that for every set T of size at most k, $\mathrm{cc}_{G+T}(b) > \mathrm{cc}_G(a)$. This is sufficient to complete the proof, since it follows from monotonicity that $\mathrm{cc}_{G+T}(a) \leq \mathrm{cc}_G(a)$. Toward this end, let T^* be a set of (at most) k edge additions such that for all sets $T \subseteq V^2 \setminus E$ of size at most k, $\mathrm{cc}_{G+T^*}(b) \leq \mathrm{cc}_{G+T}(b)$. We now state a useful structural claim, which will allow us to assume that all edges in T^* connect b to set vertices.

Claim 4.0.1. *There exists a set $T \subseteq V^2 \setminus E$ of size at most k such that every edge in T is incident on b, all other endpoints of edges in T are set vertices, and $\mathrm{cc}_{G+T}(b) \leq \mathrm{cc}_{G+T^*}(b)$.*

The condition that all edges are incident on b follows from an exchange argument, making use of the fact that (by the triangle inequality) an edge xy appears along shortest paths from b in at most one orientation, i.e., either x before y or y before x. The condition that all other endpoints are set vertices then follows from a second exchange argument, this time leveraging the particular structure of our construction. We defer the details to Appendix A.0.1.

Claim 4.0.1 allows us to give a useful lower bound on $\mathrm{cc}_{G+T^*}(b)$. We use the fact that the sets associated with the endpoints of edges in T^* are (by hypothesis) *not* a cover of U to conclude that there exists at least one element vertex v_i with $\mathrm{d}_{G+T^*}(b, v_i) = 3$. Then,

$$\begin{aligned}\mathrm{cc}_{G+T^*}(b) &\geq 2X + 2 + 2m - k + 2(n-1) + 3 = 2(n+k) + 2 + 2m - k + 2(n-1) + 3 \\ &= 4n + 2m + k + 3 > 4n + 2m + k + 2 = \mathrm{cc}_G(a),\end{aligned}$$

as desired. □

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, that CLOSENESS RATIO IMPROVEMENT is NP-hard for every $\tau > \frac{1}{2}$.

Theorem 4.0.2. CLOSENESS RATIO IMPROVEMENT *is* NP-*hard for all* $\tau \in (\frac{1}{2}, 1)$.

Proof Sketch. Let $\tau \in (\frac{1}{2}, 1)$. Consider an instance of SET COVER with m sets, n elements, and $k \in \mathbb{N}$. We first claim that we may assume $\frac{2m+4n+k}{1+2m+4n+k} \geq \tau^2$. Otherwise, we create an equivalent instance of SET COVER with $\lceil \frac{\tau^2}{4n(1-\tau^2)} \rceil$ copies of each element, giving each copy the same set memberships as the original element.[1] Then, letting n' denote the number of elements in our equivalent instance, we have

$$\begin{aligned} n' &\geq \frac{\tau^2}{4(1-\tau^2)} \\ 4n'(1-\tau^2) &\geq \tau^2 > \tau^2 + 2m(\tau^2-1) + k(\tau^2-1) \\ \frac{2m+4n'+k}{1+2m+4n'+k} &\geq \tau^2, \end{aligned}$$

so our assumption is safe. We now repeat the same construction given by Theorem 4.0.1 and depicted in Figure 4.1. This time, we let the number X of vertices in the independent set adjacent to a be an integer in the interval given by the following claim. We prove in Appendix A.0.2 that a suitable value always exists.

Claim 4.0.2. *For every $m, n, k \in \mathbb{N}$ and $\tau \in \left(\frac{1}{2}, 1\right)$, there exists a natural number X in the interval $\left(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}, \frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}\right]$.*

If there exists a set cover C of size k, then we add k edges from b to the set vertices corresponding to C. In the analysis, some care must be taken to handle the case in which these edge additions cause b's closeness centrality to become lower than that of a. For the other direction, we use our selected parameter values to show that for every set T of k edge additions, $\text{cc}_{G+T}(b) > \frac{1}{\tau} \cdot \text{cc}_G(a) \geq \frac{1}{\tau} \cdot \text{cc}_{G+T}(a)$. We defer the arithmetic to Appendix A.0.3. □

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?

[1] Conceptually, one may also think of this procedure as the introduction of $\lceil \frac{\tau^2}{4n(1-\tau^2)} \rceil$ twins for each element vertex in our constructed Closeness Ratio Improvement instance.

5

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

We now present a simple observation: if $ab \in E$, then for all vertices v, $|\mathrm{d}_G(a, v) - \mathrm{d}_G(b, v)| \le 1$. We can then use the triangle inequality to derive Observation 2, but we will also need the following observation in our proof.

Observation 1. *For all real numbers w, p, q with $0 \le w < p < q$, $\frac{p+w}{q+w} \ge \frac{p}{q}$.*

Proof. Assume toward a contradiction that $\frac{p+w}{q+w} < \frac{p}{q}$. Then $pq + qw < pq + pw$, implying that $q < p$, a contradiction. □

And now we can state the desired observation about the closeness ratio of a and b when $ab \in E$.

Observation 2. *If a and b are adjacent, then* $\text{cc-ratio}_G(a, b) \ge \frac{1}{2}$.

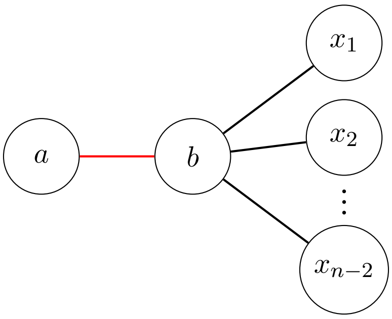

Fig. 5.1.: The worst possible closeness ratio when a and b are adjacent.

Proof. Without loss of generality, assume that $\mathrm{cc}_G(b) \leq \mathrm{cc}_G(a)$. Using the triangle inequality, if a and b are adjacent, then for any vertex v, $\mathrm{d}_G(a, v) \leq \mathrm{d}_G(a, b) + \mathrm{d}_G(b, v) = 1 + \mathrm{d}_G(b, v)$. It follows that

$$\begin{aligned}
\mathrm{cc}_G(a) &= \mathrm{d}_G(a,a) + \mathrm{d}_G(a,b) + \sum_{v \in V \setminus \{a,b\}} \mathrm{d}_G(a,v) \\
&= \mathrm{d}_G(b,b) + \mathrm{d}_G(a,b) + \sum_{v \in V \setminus \{a,b\}} \mathrm{d}_G(a,v) \\
&\leq \mathrm{d}_G(b,b) + \mathrm{d}_G(a,b) + \sum_{v \in V \setminus \{a,b\}} (1 + \mathrm{d}_G(b,v)) \\
&= \mathrm{d}_G(b,b) + \mathrm{d}_G(a,b) + \sum_{v \in V \setminus \{a,b\}} \mathrm{d}_G(b,v) + n - 2 \\
&= \mathrm{cc}_G(b) + n - 2.
\end{aligned}$$

Furthermore, it is trivially true that $\mathrm{cc}_G(b) \geq n - 1$. Thus, we may write $\mathrm{cc}_G(b) = n - 1 + x$ for some non-negative x. Then, making use of Observation 1,

$$\begin{aligned}
\text{cc-ratio}_G(a,b) = \frac{\mathrm{cc}_G(b)}{\mathrm{cc}_G(a)} \geq \frac{\mathrm{cc}_G(b)}{\mathrm{cc}_G(b) + (n-2)} &= \frac{(n-1) + x}{(n-1) + x + (n-2)} \\
&\geq \frac{n-1}{2(n-1) - 1} > \frac{1}{2}.
\end{aligned}$$

□

Recalling that no ratio greater than 1 is possible, a trivial $\frac{1}{2}$-approximation for Closeness Ratio Improvement follows: add the edge ab if it is not already present. This is also *no better* than a $\frac{1}{2}$-approximation. Given $k \geq n - 2$, an optimal algorithm achieves a ratio of 1 for the example in Figure 5.1, but the edge ab alone achieves only $\frac{1}{2} + o(1)$.

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.

6

Intuition-Building Examples

In this section, we examine three strategies for solving Closeness Ratio Improvement and demonstrate, using Figure 6.1 and Figure 6.2, that none of them yield good approximations. In Figure 6.1, the dashed edges e_1, e_2, and e_3 represent candidate edges that have not yet been added. The squiggly lines denote paths of length d, and IS_X is an independent set of X vertices, each connected to the below adjoining vertex. When calculating the closeness centrality of vertices a and b, we focus exclusively on the distances from a and b to IS_X. By choosing the parameters X and d to be sufficiently large, we can always ensure that the total distance from a and b to vertices in IS_X dominates the sum of distances to all other vertices. As a result, the contribution of vertices outside IS_X becomes asymptotically negligible when computing closeness centrality. This simplifying assumption will be used throughout the examples in this section.

Much of the existing work on centrality maximization relies on the centrality objective function being monotone and submodular (Crescenzi et al. 2016; Bergamini et al.

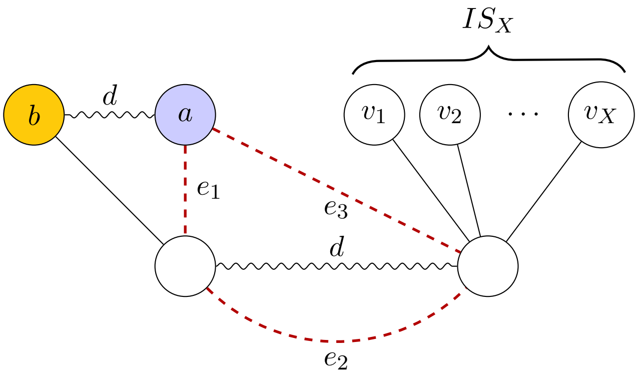

Fig. 6.1.: The construction used for counterexamples in Chapter 6. Dashed edges e_1, e_2, e_3 are potential additions, squiggly lines are paths of length d, and IS_X denotes an independent set of size X.

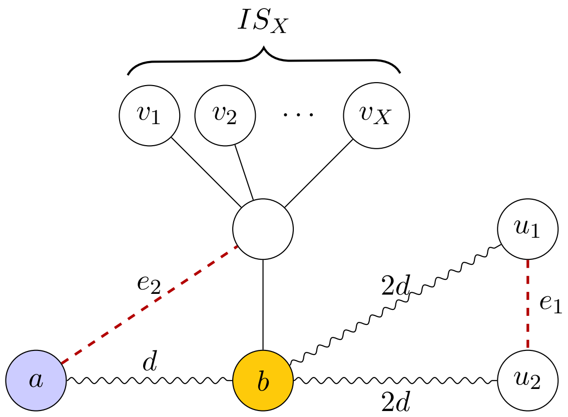

Fig. 6.2.: The construction used to demonstrate that the diameter-minimizing edge may perform arbitrarily badly compared to an optimal solution for Closeness Ratio Improvement. Dashed edges e_1, e_2 are potential additions, squiggly lines are paths of length denoted next to the edge, and IS_X denotes an independent set of size X.

2018; Medya et al. 2018), but our objective satisfies neither property. Observe that cc-ratio$_G(a, b) = \frac{d+2}{2d+2}$. Consider possible edge additions e_1 and e_2 and resulting values: cc-ratio$_{G+e_1}(a, b) = 1$, cc-ratio$_{G+e_2}(a, b) = \frac{3}{d+3}$, and cc-ratio$_{G+e_1+e_2}(a, b) = 1$. Certainly cc-ratio$_G(a, b) >$ cc-ratio$_{G+e_2}(a, b)$ shows that our objective is not monotone. We also have that, cc-ratio$_{G+e_1}(a, b) -$ cc-ratio$_G(a, b) = 1 - \frac{d+2}{2d+2} = \frac{d}{2d+2}$, while cc-ratio$_{G+e_2+e_1}(a, b) -$ cc-ratio$_{G+e_2}(a, b) = 1 - \frac{3}{d+3} = \frac{d}{d+3}$. The marginal gain from adding e_1 to the graph is larger after adding e_2, and thus our objective is not submodular.

This example also illustrates that the task of equalizing the centrality of two vertices is not necessarily the same as greedily improving the less central of the two (until they are equal). Given that a is the less central vertex in G, if we want to improve cc$_G(a)$ as much as possible with $k = 1$ in Figure 6.1, we would add e_3 to the graph. However, cc-ratio$_{G+e_3}(a, b) = \frac{2}{d+2}$, while cc-ratio$_{G+e_1}(a, b) = \frac{d+2}{d+2} = 1$. As we can increase d to make cc-ratio$_{G+e_3}(a, b)$ arbitrarily close to 0, minimizing the graph's diameter is unlikely to lead to an approximation, as this edge addition performs arbitrarily badly compared to e_1.

Finally, we use Figure 6.2 to observe that our problem diverges from the well-researched problem of adding edges to a graph to minimize its diameter (Demaine and Zadimoghaddam 2010; Li et al. 1991; Adriaens and Gionis 2022; Bilò et al. 2012). Before adding any edges, G has diameter $4d$ (the distance from u_1 to u_2). The closeness of b is around $2X$, while the closeness of a is $(d + 2)X$, giving us a ratio of $\frac{2}{d+2}$. If we add the best diameter minimizing edge, e_1 (which gets us a diameter of $3d$), we still have a ratio of $\frac{2}{d+2}$. However, if we add the edge e_2, 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.

In conclusion, three natural strategies — greedily maximizing the objective, minimizing the diameter of the graph, and greedily improving the less central vertex — all fail to yield approximations for Closeness Ratio Improvement, validating the novelty of this work.

7

Conclusion

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, in this paper we aimed 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 examined the related network-editing problem of DIAMETER MINIMIZATION (in case it may inform algorithmic approaches for our problem) in addition to more definition-based network fairness literature. We then prove that our problem, CLOSENESS RATIO IMPROVEMENT, is NP-hard for any target ratios in $\left(\frac{1}{2}, 1\right]$, 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 all perform arbitrarily badly compared to an optimal solution.

Crane et al. 2025 builds on the results from this paper, and leverages the structure of the more central vertex's neighborhood to achieve a $\frac{6}{11}$-approximation for CLOSENESS RATIO IMPROVEMENT. Additionally, they prove that CLOSENESS RATIO IMPROVEMENT is inapproximable within any factor $> \frac{5e}{5e+1-\varepsilon} \approx .932$.

Motivated by the broader goal of promoting fairness across an entire network, we now propose two new problems that extend the scope of CLOSENESS RATIO IMPROVEMENT. First, given two groups of vertices, the goal is to add at most k edges to the graph to maximize the ratio between the least central node in each group. The idea is that if the least advantaged member of each group becomes equally central, then the groups as a whole are similarly advantaged. In the following, we write cc-worst$_G(A)$ for $\arg\max_{a \in A} \mathrm{cc}_G(a)$.

GROUP CLOSENESS RATIO IMPROVEMENT

Input: A graph $G = (V, E)$, two groups $A, B \subseteq V$, $A \cap B = \emptyset$, and $k \in \mathbb{N}$.

Problem: Find $S \subseteq V^2 \setminus E$, $|S| \leq k$, maximizing cc-ratio$_{G+S}$(cc-worst$_{G+S}(A)$, cc-worst$_{G+S}(B)$).

This is a generalization of CLOSENESS RATIO IMPROVEMENT, but it seems that new insights will be needed to tackle it. For example, if $|A|, |B|$ are unbounded with respect to k, then there is not a clear way to lift the trivial approach outlined in Chapter 5. One might also consider optimizing the ratio of mean or best centralities within each group; the minimax formulation presented above is common in the fairness literature, usually with the justification that it focuses on aiding the most disadvantaged individuals.

We also propose the problem of adding edges to a graph to maximize the worst pairwise closeness ratio. A ratio close to 1 indicates that all nodes are similarly advantaged, a more equitable outcome than that offered by CLOSENESS RATIO IMPROVEMENT, which focuses on two specified nodes.

ALL-PAIRS CLOSENESS RATIO IMPROVEMENT

Input: A graph $G = (V, E)$ and $k \in \mathbb{N}$.

Problem: Find $S \subseteq V^2 \setminus E$, $|S| \leq k$, maximizing $\frac{\min_{u \in V} \mathrm{cc}_{G+S}(u)}{\max_{v \in V} \mathrm{cc}_{G+S}(v)}$.

Finally, we point out that while many algorithms focus on changing a graph to *improve* some property of a specific node, this thesis was interested in studying how to change a graph so that certain properties are *equalized* across all nodes. Future work could also explore this framework with different measures, such as betweenness, eccentricity, or degree.

A

Deferred Proofs From Chapter 4

A.0.1 Proof of Claim 4.0.1

Claim 4.0.1. *There exists a set $T \subseteq V^2 \setminus E$ of size at most k such that every edge in T is incident on b, all other endpoints of edges in T are set vertices, and* $\mathrm{cc}_{G+T}(b) \leq \mathrm{cc}_{G+T^*}(b)$.

Proof. Consider T^*, the set of at most k edges which optimally improves the closeness centrality of b. Create the set T' as follows: for each edge in T^* that is incident on b, include it in T'. For each edge $xy \in T^*$ that is not incident on b, suppose without loss of generality that $\mathrm{d}_{G+T^*}(b,x) \leq \mathrm{d}_{G+T^*}(b,y)$, and then include the edge by in T'. Now $|T'| \leq k$ and all edges in T are incident on b.

For an arbitrary $v \in V$, either the shortest path from b to v in $G + T^*$ used an edge in T^* not incident on b, or it did not. In the first case, let xy be the last edge in T^* not incident on b used by the shortest path from b to v in $G + T^*$. Then we know that the edge by exists in $G + T'$, by construction. Furthermore, as xy is the last edge in T^* on the shortest path from b to v, with y as the further endpoint from b, $\mathrm{d}_G(y,v) = \mathrm{d}_{G+T^*}(y,v) = \mathrm{d}_{G+T'}(y,v)$. Thus,

$$\mathrm{d}_{G+T'}(b,v) \leq \mathrm{d}_{G+T'}(b,y) + \mathrm{d}_{G+T'}(y,v) = 1 + \mathrm{d}_{G+T'}(y,v)$$

$$\leq \mathrm{d}_{G+T^*}(b,y) + \mathrm{d}_{G+T^*}(y,v) = \mathrm{d}_{G+T^*}(b,v)$$

In the second case, if the shortest path from b to v in $G + T^*$ did not use an edge in T^* not incident on b, then all the edges on the path will still exist in $G + T'$, by construction. Thus $\mathrm{d}_{G+T'}(b,v) \leq \mathrm{d}_{G+T^*}(b,v)$.

So for all vertices v, $\mathrm{d}_{G+T'}(b,v) \leq \mathrm{d}_{G+T^*}(b,v)$, implying that $\mathrm{cc}_{G+T'}(b) \leq \mathrm{cc}_{G+T^*}(b)$. We'll now use a second exchange argument on T' to construct a set of at most k edges T such that every edge in T is incident on b, all other endpoints of edges in T are set vertices, and $\mathrm{cc}_{G+T}(b) \leq \mathrm{cc}_{G+T'}(b) \leq \mathrm{cc}_{G+T^*}(b)$.

For each edge in T' not incident on a set vertex, either the edge is incident on a vertex z_i in the independent set, or it is incident on an element vertex v_i. In the first case, bz_i reduces the closeness centrality of b by exactly 1, as it has decreased $\mathrm{d}_{G+T'}(b, z_i)$ from 2 to 1, but not reduced any other distances. In the second case, bv_i reduces the closeness centrality of b by exactly 2, as it has decreased $\mathrm{d}_{G+T'}(b, v_i)$ from 3 to 1, but not reduced any other distances.

As we assume there is no set cover, for any set S of at most k edges, there must exist some element vertex v_i with no edges in S incident on it or the vertex of any set containing it. So given T', there exists such a v_i, and then connect b to a set vertex s_j, where $v_i \in S_j$. Then $\mathrm{d}_{G+T'}(s_j, b) = 2$ and $\mathrm{d}_{G+T'}(v_i, b) = 3$, while $\mathrm{d}_{G+T'+\{bs_j\}}(s_j, b) = 1$ and $\mathrm{d}_{G+T'+\{bs_j\}}(v_i, b) = 2$. Therefore, we can always find an edge bs_j which reduces the closeness centrality of b by at least 2 when added to the graph.

So construct T as follows: for each edge in T' that is incident on a set vertex, include it. For each edge in T' that is not incident on a set vertex, include the edge bs_j in T, where s_j is a set vertex with some adjacent element vertex uncovered (we've shown that a suitable s_j always exists, as long as our edge set is at most k edges and there is no set cover). As argued above, each edge substitution does not increase the closeness centrality achieved by the edge set at all, and thus $\mathrm{cc}_{G+T}(b) \leq \mathrm{cc}_{G+T'}(b) \leq \mathrm{cc}_{G+T^*}(b)$, as desired. □

A.0.2 Proof of Claim 4.0.2

Claim 4.0.2. *For every $m, n, k \in \mathbb{N}$ and $\tau \in \left(\frac{1}{2}, 1\right)$, there exists a natural number X in the interval $\left(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}, \frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}\right]$.*

Proof. Taking the difference of the bounds of the interval, we get that the length of the interval is $\frac{\tau}{2\tau-1}$. If $\tau < 1$, then $-\tau > -1$ and thus $\tau > 2\tau - 1$. Thus the length of the interval is greater than 1 when $\tau \in (\frac{1}{2}, 1)$, so there must exist some integer X

within this interval. Furthermore, for $m, n, k \in \mathbb{N}$ and $\tau \in (\frac{1}{2}, 1)$, the bounds of this interval are both positive, so X must be a natural number, as desired. □

A.0.3 Remainder of the Proof of Theorem 4.0.2

Theorem 4.0.2. Closeness Ratio Improvement *is* NP-*hard for all* $\tau \in (\frac{1}{2}, 1)$.

Proof Conclusion. See Chapter 4 for the construction and beginning of the analysis. All that remains is to show that our constructed Closeness Ratio Improvement instance is a yes-instance if and only if the Set Cover instance from which we reduce is a yes-instance.

For the forward direction, suppose that there exists a cover $C \subseteq \{S_j\}_{j=1}^{m}$ of size k. Again, let $T = \{bs_j \colon S_j \in C\}$, the set of edges from b to the set vertices corresponding to sets in the cover C. Now, $\text{cc}_{G+T}(a) = X + 2 + 2m + 3n$ and $\text{cc}_{G+T}(b) = 2X + 2 + 2m - k + 2n$. We want to show that $\text{cc-ratio}_{G+T}(a, b) \geq \tau$, but we must still ensure that $\text{cc-ratio}_{G+T}(a, b) \leq 1$, or else we violate our min/max definition of closeness ratio. Note that this was not a concern for $\tau = 1$, as we showed our ratio was exactly 1 when there was a set cover. Thus we consider the case where $\text{cc}_{G+T}(b) \geq \text{cc}_{G+T}(a)$ and the case where $\text{cc}_{G+T}(b) < \text{cc}_{G+T}(a)$, and show that in both cases $\text{cc-ratio}_{G+T}(a, b) \geq \tau$.

If $\text{cc}_{G+T}(b) \geq \text{cc}_{G+T}(a)$, then $\text{cc-ratio}_{G+T}(a, b) = \frac{X+2+2m+3n}{2X+2+2m-k+2n}$. Observe that this ratio is greater than $\frac{1}{2}$, and as $X \to \infty$, $\text{cc-ratio}_{G+T}(a, b) \to \frac{1}{2}$. Thus $\text{cc-ratio}_{G+T}(a, b)$ is a decreasing function of X and using Claim 4.0.2, we have

$$\text{cc-ratio}_{G+T}(a, b) \geq \frac{(\frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}) + 2 + 2m + 3n}{2(\frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}) + 2 + 2m - k + 2n}$$

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

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

Alternatively, if $\text{cc}_{G+T}(b) < \text{cc}_{G+T}(a)$, then $\text{cc-ratio}_{G+T}(a, b) = \frac{2X+2+2m-k+2n}{X+2+2m+3n}$.

Observe that this ratio is now an *increasing* function of X and using Claim 4.0.2 we have,

$$\text{cc-ratio}_{G+T}(a,b) > \frac{2(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}) + 2 + 2m - k + 2n}{(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}) + 2 + 2m + 3n}$$

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

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

$$= \frac{4+4m+6n-2\tau-2-2m+k-2n}{4\tau+4m\tau+6n\tau-2\tau-2m\tau+k\tau-2n\tau-\tau} = \frac{(2-2\tau)+2m+4n+k}{\tau+2m\tau+4n\tau+k\tau}.$$

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

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

We showed in our construction in Chapter 4 that we may assume $\frac{2m+4n+k}{1+2m+4n+k} \geq \tau^2$. Thus we have that,

$$\text{cc-ratio}_{G+T}(a,b) > \frac{1}{\tau}(\frac{2m+4n+k}{1+2m+4n+k}) \geq \frac{1}{\tau}(\tau^2) = \tau.$$

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

For the reverse direction, suppose that there is no collection of k sets which covers U. We must prove that there is no set $T \subseteq V^2 \setminus E$ of size at most k such that cc-ratio$_{G+T}(a,b) \geq \tau$. Let T^* be a set of (at most) k edge additions such that for all sets $T \subseteq V^2 \setminus E$ of size at most k, $\text{cc}_{G+T^*}(b) \leq \text{cc}_{G+T}(b)$. Using the same analysis from the proof of Theorem 4.0.1, $\text{cc}_{G+T^*}(b) \geq 2X + 2 + 2m - k + 2n + 1$, while $\text{cc}_{G+T^*}(a) \leq X + 2 + 2m + 3n$. Then, cc-ratio$_{G+T^*}(a,b) \leq \frac{X+2+2m+3n}{2X+2+2m-k+2n+1}$. Note that our choice of X (Claim 4.0.2) guarantees that $\text{cc}_{G+T^*}(a) \leq \text{cc}_{G+T^*}(b)$, as we are about to show. This ratio is greater than $\frac{1}{2}$, and as $X \to \infty$, cc-ratio$_{G+T^*}(a,b) \to \frac{1}{2}$. Thus cc-ratio$_{G+T^*}(a,b)$ is a decreasing function of X and we have,

$$\text{cc-ratio}_{G+T^*}(a,b) < \frac{(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}) + 2 + 2m + 3n}{2(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}) + 2 + 2m - k + 2n + 1}$$

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

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

□

References

Adriaens, Florian and Aristides Gionis (2022). “Diameter Minimization by Shortcutting with Degree Constraints”. In: DOI: 10.48550/arXiv.2209.00370 (cit. on pp. 7, 8, 26, 34).

Bashardoust, Ashkan, Sorelle Friedler, Carlos Scheidegger, Blair D. Sullivan, and Suresh Venkatasubramanian (2023). “Reducing Access Disparities in Networks using Edge Augmentation”. In: *FAccT ’23*, pp. 1635–1651 (cit. on pp. 4, 19, 21).

Bergamini, Elisabetta, Pierluigi Crescenzi, Gianlorenzo D’Angelo, Henning Meyerhenke, Lorenzo Severini, and Yllka Velaj (2018). “Improving the Betweenness Centrality of a Node by Adding Links”. In: *Journal of Experimental Algorithmics* 23, pp. 1–32 (cit. on p. 33).

Bhaskara, Aditya, Alex Crane, Shweta Jain, Md Mumtahin Habib Ullah Mazumder, Blair D. Sullivan, and Prasanth Yalamanchili (2024). “Optimizing Information Access in Networks via Edge Augmentation”. In: arXiv: 2407.02624 [cs.DS]. URL: https://arxiv.org/abs/2407.02624 (cit. on p. 4).

Bilò, Davide, Luciano Gualà, and Guido Proietti (2012). “Improved approximability and non-approximability results for graph diameter decreasing problems”. In: *Theoretical Computer Science* 417, pp. 12–22 (cit. on pp. 7, 34).

boyd, danah, Karen Levy, and Alice Marwick (2014). “The Networked Nature of Algorithmic Discrimination”. In: *Data and Discrimination: Collected Essays*, pp. 43–57 (cit. on p. 2).

Crane, Alex, Sorelle A. Friedler, Mihir Patel, and Blair D. Sullivan (2025). *Equalizing Closeness Centralities via Edge Additions*. arXiv: 2505.06222 [cs.DS]. URL: https://arxiv.org/abs/2505.06222 (cit. on pp. 25, 36).

Crescenzi, Pierluigi, Gianlorenzo D’Angelo, Lorenzo Severini, and Yllka Velaj (2016). “Greedily Improving Our Own Closeness Centrality in a Network”. In: *ACM Transcations on Knowledge Discovery from Data* 11.1, pp. 1–32 (cit. on p. 33).

Demaine, Erik and Morteza Zadimoghaddam (2010). “Minimizing the Diameter of a Network using Shortcut Edges”. In: *Algorithm Theory - SWAT 2010* 6139, pp. 420–431 (cit. on p. 34).

Fish, Benjamin, Ashkan Bashardoust, Danah Boyd, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian (2019). “Gaps in Information Access in Social Networks”. In: DOI: 10.1145/3308558.3313680 (cit. on pp. 19, 21).

Gonzalez, Teofilo F. (1985). “Clustering to minimize the maximum intercluster distance”. In: *Theoretical Computer Science* 38, pp. 293–306 (cit. on p. 11).

Karp, Richard (1972). “Reducibility among Combinatorial Problems”. In: *Complexity of Computer Computations*, pp. 85–103 (cit. on p. 7).

Li, Chung-Lun, S. Thomas McCormick, and David Simchi-Levi (1991). “On the Minimum-Cardinality-Bounded-Diameter and the Bounded-Cardinality-Minimum-Diameter Edge Addition Problems”. In: *Operations Research Letters* 11 (5), pp. 303–308 (cit. on pp. 7, 13–15, 34).

Medya, Sourav, Arlei Silva, Ambuj Singh, Prithwish Basu, and Ananthram Swami (2018). "Group Centrality Maximization via Network Design". In: pp. 126–134. ISBN: 978-1-61197-532-1 (cit. on p. 34).

Meyerson, Adam and Brian Tagiku (2009). "Minimizing Average Shortest Path Distances via Shortcut Edge Addition". In: *APPROX 2009*, pp. 272–285 (cit. on p. 15).

Yeh, Min-Hsuan, Blossom Metevier, Austin Hoag, and Philip Thomas (2024). "Analyzing the Relationship Between Difference and Ratio-Based Fairness Metrics". In: *The 2024 ACM Conference on Fairness, Accountability and Transparency*, pp. 518–528 (cit. on p. 24).