A Foundational Perspective for Partitional Clustering on Networks

Derya Ipek Eroglu[a], Cem Iyigun[b]

[a] *Computing Sciences, SUNY Brockport, 350 New Campus Drive, Brockport, 14420, New York, USA*
[b] *Industrial Engineering, Middle East Technical University, Üniversiteler Mahallesi, Dumlupınar Bulvarı No:1, 06800, Ankara, Turkey*

Abstract

This study presents a theoretical analysis of partitional clustering on networks. Different versions of the problem are studied considering different assignment schemes (hard and soft) and different objective functions. Cluster centers are not restricted to only the set of nodes, it is assumed that centers can also be at the edges of the network. Four clustering problems are investigated namely P-Median (PMP), Sum of Squares Clustering (SSC) under hard assignment, as well as the Probabilistic Distance Clustering (PDC) and Fuzzy C-Means Clustering (FCM) problems under soft assignment. Through mathematical analysis, we establish new structural properties of these problems, such as the significance of assignment bottleneck points and the role of vertex-restricted solutions in determining optimal cluster centers. Our findings demonstrate that, while SSC and FCM problems allow optimal centers to be positioned along edges, PMP and PDC inherently favor vertex placement, leading to insights into clustering behavior on networks. These results open avenues for the development of efficient heuristics and metaheuristics that leverage these properties, with potential applications in facility location, network design, and data clustering on networks.

Keywords: Network Clustering, Network Location Problems, Center-based Clustering, P-median Problem, Probabilistic Distance Clustering, Sum of Squares Clustering, Fuzzy C-means Clustering, K-means Clustering

1. Introduction

Clustering is an unsupervised learning method that groups similar data points and separates dissimilar ones based on a defined distance metric [1]. It is widely used as an exploratory analysis technique [2], uncovering underlying patterns and structures within complex datasets [3]. Clustering has been extensively studied across various domains, with existing literature primarily categorized into two main types: partitional clustering and hierarchical clustering. Partitional clustering partitions data into distinct, non-overlapping subsets, while hierarchical clustering organizes data into nested structures based on hierarchical relationships.

While traditional clustering methods are well-developed for continuous spaces, many real-world datasets are inherently relational, best represented as networks. This necessitates clustering methods that account for graph connectivity and discrete relationships. Consequently, graph clustering has gained increasing attention, with applications in social networks, biological systems, and transportation networks [4]. Recent advances in clustering have expanded beyond standard partitioning techniques to incorporate more flexible models, such as soft clustering [5] and density-controlled clustering [6], making clustering more adaptable to real-world constraints.

Despite these advancements, many existing graph clustering approaches prioritize partitioning strategies rather than optimizing cluster center locations, limiting their applicability in certain theoretical and practical settings [5]. However, location theory—a field concerned with optimal facility placement in networks—provides a natural framework to extend clustering models by incorporating optimization techniques used in network design and logistics. Recent work on p-median and facility location models [7, 8] highlights the importance of optimizing cluster centers, particularly in applications such as wireless sensor networks, urban facility placement, and transportation planning. Furthermore, clustering methods that integrate traffic-aware constraints [7] and user-specified density parameters [6] have demonstrated the growing need for models that balance theoretical rigor with practical applicability.

In this paper, we propose a unified theoretical framework for analyzing partitional clustering in networks, focusing on two key aspects: assignment schemes and objective functions. We examine both hard clustering, where each data point is assigned to a single cluster, and soft clustering, which allows fractional assignments and enables partial memberships. Our focus on soft clustering connects directly to recent advances which unravels additional benefits coming with using soft assignment [9]. Additionally, we explore clustering models that minimize either the sum of distances or the sum of squared distances between nodes and their assigned cluster centers. These formulations align with recent work on density-aware clustering [6] and p-median problem [7], both of which highlight the importance of the problems of interest. By establishing theoretical connections between clustering methodologies and location theory, we analyze the behavior of optimal solutions in networks, providing fundamental insights into the structural properties of these models. This theoretical foundation bridges recent advances in graph clustering and facility location models, offering a novel lens for understanding clustering in networks.

The remainder of this paper is structured as follows: Section 2 formally defines the clustering problems under study. Section 3 introduces the theoretical framework supporting our analysis, while Section 4 presents our key theoretical results. Finally, Sections 5 and 6 discuss our contributions and outline potential directions for future research.

2. Problem Definitions and Network Setting

Table 1: Categorization of Clustering Problems by Assignment Type and Objective Function with Mathematical Formulations

		Assignment Type	
		Hard Assignment	**Soft Assignment**
Objective Function	**Distance**	**K-median** $\min \sum_{k\in K} \sum_{i\in C_k} \{d(x_i, c_k)\}$	**PD-Clustering** $\min \sum_{i\in I} \sum_{k\in K} p_{ik}^2 d(x_i, c_k)$
	Squared Distance	**K-Means** $\min \sum_{k\in K} \sum_{i\in C_k} \{d(x_i, c_k)^2\}$	**Fuzzy C-Means** $\min \sum_{i\in I} \sum_{k\in K} p_{ik}^m d(x_i, c_k)^2$

Clustering has been extensively studied in planar setting, providing crucial insights into data partitioning strategies. In this paper, we focus on four clustering models, distinguished by two primary factors: the assignment type (hard vs. soft) and the objective function (sum of distances vs. sum of squared distances). Table 2 summarizes these variations and their corresponding objective functions.

Let I be the set of data points and K the set of clusters. In a the planar setting, each data point is represented by coordinates $x_i \in \mathbb{R}^d$, while cluster centers are denoted $c_k \in \mathbb{R}^d$. The function $d(x_i, c_k)$ typically refers to the Euclidean distance between the data points and cluster centers.

Hard assignment. Each data point belongs to exactly one cluster. Two well-known hard-assignment algorithms are:

- **K-median**, which minimizes the sum of distances between data points and their respective cluster centers. It is known for its robustness to outliers.
- **K-Means**, which minimizes the sum of squared distances to cluster centroids, emphasizing the reduction of within-cluster variance.

Soft assignment. A data point can belong to multiple clusters with fractional memberships p_{ik}. Two prominent examples are:

- **Fuzzy C-Means (*FCM*)** [10], which uses a fuzziness parameter m. As m increases, cluster boundaries become more blurred, and membership values p_{ik} approach a uniform distribution [11].
- **Probabilistic Distance Clustering (PD-Clustering)** [12], which also allows fractional memberships but minimizes the sum of distances, aligning it more closely with K-Medoids in terms of the objective function.

Although these formulations provide a strong theoretical foundation in planar contexts, many real-world datasets are naturally represented as networks. In such graph-based settings, the vertices represent the data points, and the edges define pairwise relationships. Hence, clustering on networks requires shortest path distances rather than Euclidean distances. As can be inferred from the comprehensive surveys [5, 13], many network clustering approaches focus on partitioning the graph but do not explicitly address the problem of optimally locating cluster centers.

A key exception is the well-studied P-Median Problem, an example of hard assignment on a network, which seeks to minimize the sum of shortest path distances between each **node** and its assigned **center**. Hakimi's pioneering results [14, 15] showed that optimal centers for the P-Median problem lie at vertices. Levy [16] generalized this result and found that the root of this behavior is the objective function concavity. Building on Hakimi's foundational insights [14, 15] and Levy's extension [16], we introduce a unified framework that not only leverages these properties but also offers new perspectives, proofs, and a comparative treatment of four partitional clustering problems considering different assignment schemes using different objective functions. These foundational insights motivate to investigate different clustering problems to analyze the behaviour of the objective functions and derive the structural properties of selecting clusters centers for each underlying problem.

To adapt the four clustering models in Table 2 to a network, we denote vertices as v_i and restrict each cluster center x_k to be on the network. The distance function $d(v_i, x_k)$ then refers to the shortest path distance between vertex v_i and the center x_k. In the sections that follow, we analyze how these four models behave in networks, highlighting both shared and distinctive structural properties.

3. Fundamentals and Background

In this section we introduce the basic notation and some concepts that will be used in the analysis of the models.

The clustering problems of interest are defined on an undirected, connected graph $\mathbf{G} = (\mathbf{E}, \mathbf{V})$, where $\mathbf{V}$ denotes the set of vertices and $\mathbf{E}$ represents the set of edges. Since $\mathbf{G}$ is connected, it follows that $|\mathbf{E}| \geq |\mathbf{V}| - 1$, ensuring that every vertex is reachable.

On this graph, we define $d(v_i, x_k)$ as the length of the **shortest path distance** between vertex v_i and cluster center x_k. Since $\mathbf{G}$ is undirected, it follows that $d(v_i, x_k) = d(x_k, v_i)$. The underlying distance measure satisfies the properties of a metric space, with Euclidean distance as the assumed metric unless otherwise stated. While alternative distance metrics can be incorporated within this framework, our analysis focuses on Euclidean metric. Table 2 summarizes the notation used throughout the paper.

By incorporating these notations, we extend the clustering models from Table 2 to net-

Table 2: Table of Notations

$\mathbf{V}$	Set of vertices
$\mathbf{X}$	Set of cluster centers
n	Number of vertices $\mathbf{V}$
$\mathbf{I}$	Index set of vertices $I = \{1, 2, ..., n\}$
v_i	Vertex i, where $i \in \mathbf{I}$
h_i	Weight of v_i, where ($h_i > 0 \ \forall i \in \mathbf{I}$)
b_i	Arc bottleneck point of v_i on the edge (v_p, v_q)
a_i	Assignment bottleneck point of v_i
p	Number of clusters $\lvert\mathbf{X}\rvert$
x_k	Location of cluster center k
$d(v_i, x_k)$	Length of the shortest path from v_i to x_k
p_{ik}	Probability of assignment of v_i to cluster k

works by adding the center location constraint and updating the distance function. We next establish two key concepts—arc bottleneck points and assignment bottleneck points—to characterize the solution behavior on networks.

3.1. Arc Bottleneck Point

Let v_i be an arbitrary vertex in $\mathbf{G}$, and let e_{pq} be an edge connecting vertices v_p and v_q with length l_e. For any point $x \in e_{pq}$, the shortest path distance from v_i to x is given by:

$$d(v_i, x) = \min\{d(v_i, v_p) + d(v_p, x), d(v_i, v_q) + d(v_q, x)\}. \tag{1}$$

This formulation implies that the shortest path to x passes through either v_p or v_q. Figure 1 illustrates this relationship. There exists a point, b_i, on edge e_{pq} at which the shortest paths via v_p and v_q are of equal length. This point, called the /textbfarc bottleneck point, represents the farthest location on e_{pq} from v_i that maintains the shortest path property. The location of b_i is determined by:

$$d(v_i, v_p) + d(v_p, b_i) = d(v_i, v_q) + d(v_q, b_i)$$

Substituting $d(v_q, b_i) = l_e - d(v_p, b_i)$, we obtain:

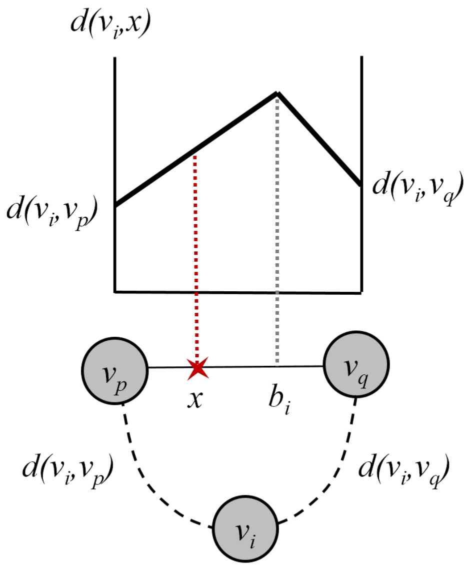

Figure 1: Distance function $d(v_i, x_)$ on the edge connecting v_p and v_q

$$
\begin{aligned}
d(v_i, v_p) + d(v_p, b_i) =& d(v_i, v_q) + l_e - d(v_p, b_i) \\
d(v_p, b_i) =& \frac{1}{2}(d(v_i, v_q) + l_e - d(v_i, v_p)),
\end{aligned}
\tag{2}
$$

The arc bottleneck point b_i has been studied in facility location problems. Before the study in [17], Hakimi used this concept in his proof in [14]. There are three cases regarding the value of $d(v_p, b_i)$:

Case 1 . If $d(v_p, b_i) \leq 0$, the shortest path from v_i to x always passes from v_q.

Case 2 . If $d(v_p, b_i) \geq l_e$, the shortest path from v_i to x always passes from v_p.

Case 3 . If $d(v_p, b_i) \in (0, l_e)$, the shortest path from v_i to x passes from:

- v_p if $d(v_p, x) <= d(v_p, b_i)$,
- v_q if $d(v_p, x) > d(v_p, b_i)$.

For cases with multiple vertices, additional bottleneck points can emerge, as illustrated in Figure 2. Since end vertices do not create bottleneck points, the number of arc bottleneck points on an edge is at most $n - 2$.

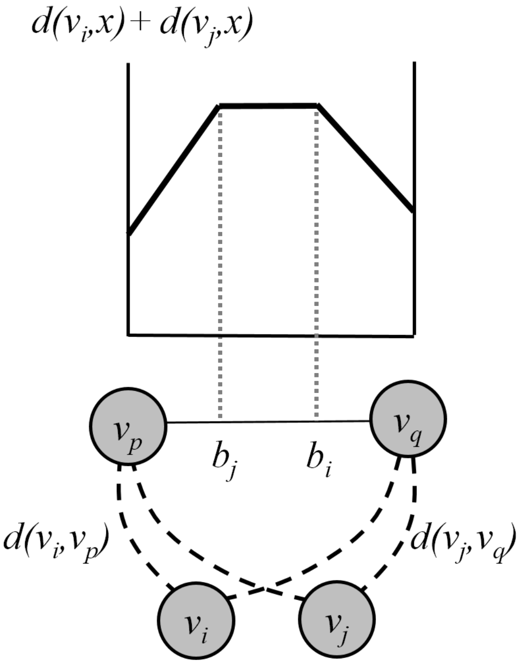

Figure 2: Distance function $d(v_i, x) + d(v_j, x)$ on the edge connecting v_p and v_q

3.2. Assignment Bottleneck Point

While arc bottleneck points capture the farthest shortest-path location on an edge, they do not account for cluster assignment changes when centers move. We introduce the assignment bottleneck point, which identifies the location where a vertex switches its nearest cluster center.

Let v_i be an arbitrary vertex in $\mathbf{G}$, and let $x_k \in e, k = 1, ..., K$ be the closest center to v_i. Assume that $x_l \in \mathbf{G}$ is the second-closest center to v_i, and that all cluster center locations except x_k remain fixed. As x_k moves along e_{pq}, its distance to v_i changes, and at a critical point, x_l may become closer than x_k. The location where this assignment shift occurs is defined as the assignment bottleneck point. Figure 3 illustrates an example where an assignment bottleneck point a_i appears. Initially, $d(v_i, x_k) = 13$ and $d(v_i, x_l) = 15$, with x_k as the closest center. As x_k moves toward v_p, its distance to v_i increases. If x_k moves more than 2 units, the assignment changes to x_l, as $d(v_i, x_k) \geq d(v_i, x_l)$.

If an arc bottleneck point exists on e_{pq}, multiple assignment bottleneck points may arise. Figure 4 illustrates a scenario where: 1. The assignment switches at a_i^q, making x_l the closest center. 2. The assignment switches back at a_i^p, reassigning x_k as the closest center.

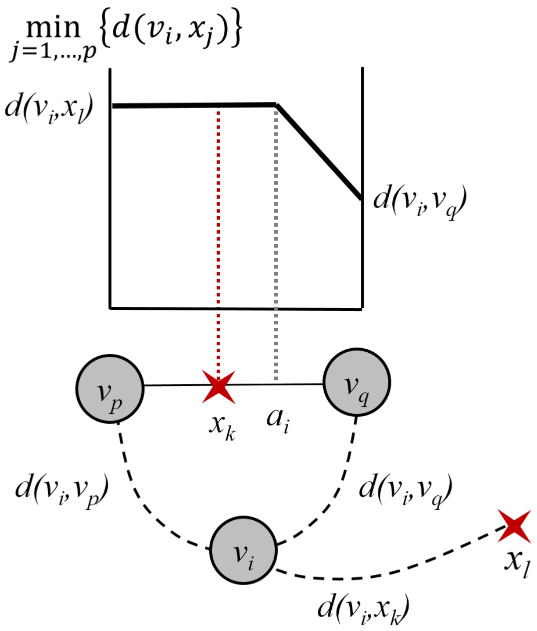

Figure 3: Assignment Bottleneck Point

Each edge can contain at most $2(n-2)+2$ assignment bottleneck points. The presence of an arc bottleneck point is necessary for an edge to have two assignment bottleneck points for a given vertex.

Having discussed the fundamental definitions for both hard and soft assignment clustering in network settings, next section focuses on the core theoretical aspects of these models. We analyze the objective function behavior of all four problems and establish the theorems that underpin our analytical results. This framework will clarify how and why optimal solutions tend to concentrate on specific locations of **G**, ultimately revealing deeper insights into the nature of cluster center placement in networks.

4. Structural Insights and Analytical Results

In this section, we conduct our analysis under hard and soft assignment schemes and different objective functions as described in Table 2.

4.1. Hard Assignment Problems

In the hard assignment case, each vertex is assigned to a single cluster, determined by its closest cluster center. Two key hard clustering problems are examined in this section:

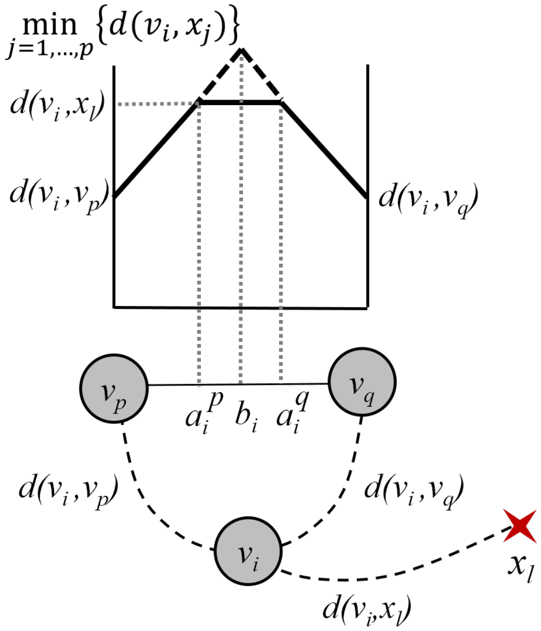

Figure 4: Assignment Bottleneck Point with Arc Bottleneck Point

the P-Median Problem, which minimizes the sum of distances between vertices and their assigned cluster centers, and the SSC, which minimizes the sum of squared distances. We begin with the P-Median Problem and extend the discussion to the SSC, analyzing their theoretical properties in a network setting.

4.1.1. P-Median Problem

The P-Median Problemis formally defined as

$$\textit{minimize}\ f(\mathbf{X}) = \sum_{i=1}^{n} h_i \min_{k=1,\dots,p}\{d(v_i, x_k)\}$$

$$\text{subject to}$$

$$x_k \in V \quad \forall\, k = 1, \dots, p,$$

where $\mathbf{X}$ represents the set of cluster centers, x_k is the decision variable for the location of center k, and h_i is a nonnegative constant weight associated with vertex v_i. Since hard assignment is imposed, each vertex is exclusively assigned to one cluster, and only the distance to its closest center contributes to the objective function. We first analyze the case with a single cluster (the 1-Median Problem) and subsequently extend the results to the general

P-Median Problem.

1-Median Problem. Suppose we have only one cluster and one cluster center will be located on $\mathbf{G}$. In that case, the formulation is

$$minimize \ f(x_c) = \sum_{i=1}^{n} h_i d(v_i, x_c) \tag{3}$$

$$\text{subject to}$$

$$x_c \in \mathbf{G}. \tag{4}$$

Consider a simple line graph with four vertices as in Figure 5. If the cluster center is located anywhere on the graph, the objective function remains constant. This implies that the optimal center location could be on either a vertex or an edge.

$$f = (a + y) + y + (l - y) + (b + l - y) = a + 2l + b,$$

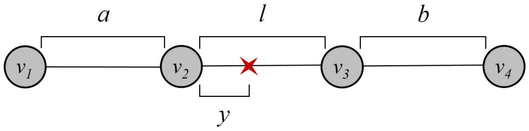

Figure 5: A Line Graph with 4 vertices

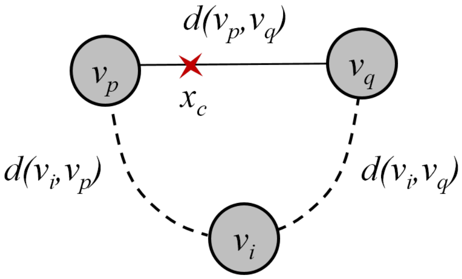

Figure 6: An illustration of a part of a graph $\mathbf{G}$

A general network is illustrated in Figure 6 in which cluster center x_c is on an edge e_{pq} connecting (v_p, v_q). In the case of this network, we may observe three different patterns of objective function component of a vertex v_i to the (3) depending on location of x_c. The

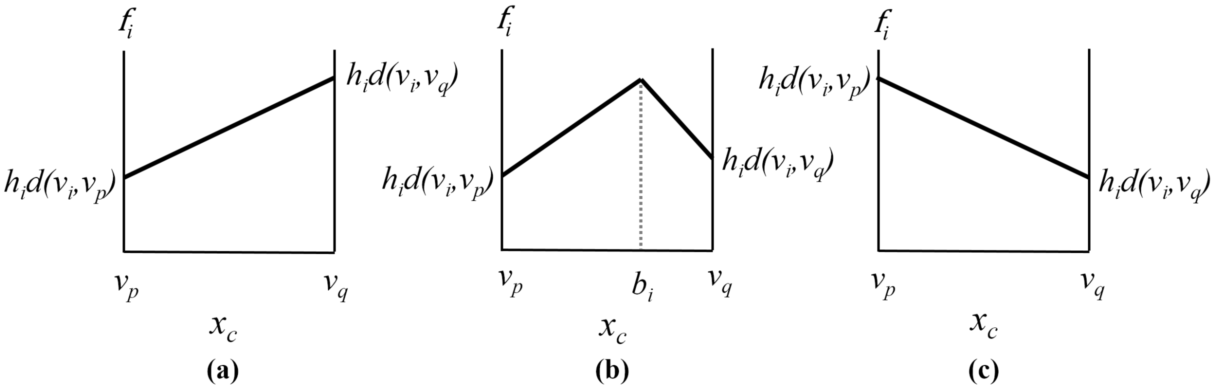

Figure 7: Objective function component for v_i (denoted as f_i) when x_c is moved along the edge (v_p, v_q)

shortest path from v_i to x_c may pass through v_p or v_q regardless of the location of x_c, which also means that there is no arc bottleneck point. If there is an arc bottleneck point on e_{pq}, the shortest path to x_c will pass from v_p or v_q depending on the location of x_c on e_{pq}. These patterns are shown in Figure 7. In the Figure, (a) is the case when shortest path to x_c passes from v_p, and (c) is the case when shortest path to x_c passes through v_q. In both cases, there is no arc bottleneck point. In (b), if x_c is in the interval $[v_p, b_i]$, shortest path to x_c passes through v_p; otherwise, shortest path to x_c passes through v_q. The reason of this behavior is the bottleneck point b_i observed. In all of these cases, it could be observed that the objective function is linear or piecewise concave.

Using this structure, Hakimi [14] proved that the optimal center locations always lie in $\mathbf{V}$. Levy [16] extended this proof by showing that this result is a consequence of the concavity of the objective function. Since the summation of concave functions remains concave, the optimal solution to the 1-Median problem is always located at a vertex.

P-Median Problem. In P-Median Problem, vertices are assigned to the clusters with the closest cluster center minimizing the p-median objective function in Table 2. For the general P-Median Problem, Hakimi [15] extended his previous proof and showed that the optimal cluster centers are always located in $\mathbf{V}$. His proof follows from decomposing the problem into p separate 1-Median problems and showing that each center must be positioned at a vertex.

An alternative approach is to analyze the behavior of the objective function along the edges. Consider a network with p cluster centers, where x_c and x_k are the closest and second-closest cluster centers to v_i, respectively. If x_c is moved along an edge (v_p, v_q), the assignment of v_i to a cluster may change. These assignment changes occur at assignment bottleneck points, as shown in Figure 9.

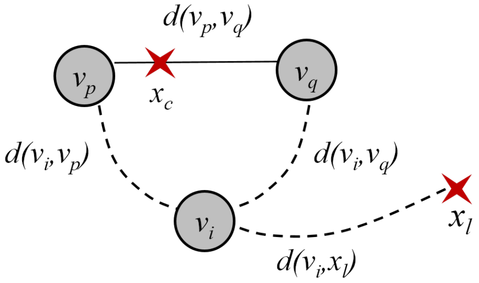

Figure 8: A visualization of a part of graph $\mathbf{G}$ with 2 closest cluster centers

Even with multiple centers, the objective function remains piecewise concave due to the presence of arc bottleneck points and assignment bottleneck points. Since the sum of concave functions is also concave, the optimal centers are always located in $\mathbf{V}$, as proved in [16].

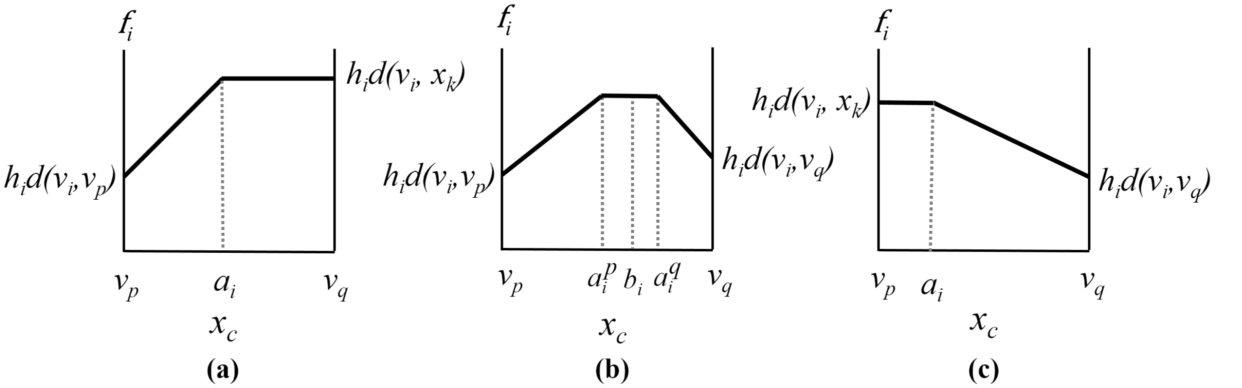

Figure 9: Objective function component for v_i (denoted as f_i) when x_c is moved along the edge (v_p, v_q) and x_k is the second closest cluster center to v_i

4.1.2. Sum of Squares Clustering (SSC) Problem on Networks

SSC problem differs from P-Median Problem in that it uses sum of squared distances in the objective function instead of sum of distances. SSC problem is defined as

$$\textit{minimize}\ f(\mathbf{X}) = \sum_{i=1}^{n} h_i \min_{k=1,\dots,p} \{d(v_i, x_k)^2\}$$

$$\text{subject to}$$

$$x_k \in G \quad \forall\, k = 1, ..., p,$$

where x_k is the decision variable representing the location of cluster center k, and h_i is a nonnegative constant weight associated with vertex v_i. Due to hard assignment constraints, each vertex is assigned to exactly one cluster.

A key difference from the P-Median Problem is that, as observed by Carrizosa et al. [18], the optimal center locations may lie not only in $\mathbf{V}$ but also in $\mathbf{E}$. Consequently, restricting cluster center locations solely to $\mathbf{V}$ could lead to suboptimal solutions. In this subsection, we analyze this property in more detail.

SSC on Networks with a Single Cluster. As in P-Median Problem, when there is only one cluster, every vertex belongs to that cluster. Consider the simple line graph with four vertices shown in Figure 5. The objective function for the SSCproblem is given by:

$$f = (a + y)^2 + y^2 + (l - y)^2 + (b + l - y)^2.$$

This function is continuous and twice differentiable. Its first and second derivatives with respect to y are:

$$\frac{df}{dy} = 8y + 2a - 2l - 2b - 2l, \tag{5}$$

$$\frac{d^2 f}{dy^2} = 8, \tag{6}$$

which shows that f is a convex function of y. (5) gives us

$$y = \frac{b + 2l - a}{4}.$$

Therefore, the optimal center location depends on the following cases:

- If $y \in (0, l)$, the optimal center is located within the edge (v_2, v_3).
- If $y = 0$, the optimal center is at vertex v_2.
- If $y = l$, the optimal center is at vertex v_3.
- If $y \in (0, -a)$, the optimal center is within the edge (v_1, v_2).
- If $y = -a$, the optimal center is at vertex v_1.
- If $y \in (l, b + l)$, the optimal center is within the edge (v_3, v_4).
- If $y = b + l$, the optimal center is at vertex v_4.

Thus, in contrast to the P-Median Problem, the optimal solution for SSC is not necessarily at a vertex but may be within an edge.

Consider a generalized line graph, as in Figure 6, where the cluster center x_c is located on an edge (v_p, v_q). The objective function is:

$$f = \sum_{i=1}^{n} h_i d(v_i, x_c)^2. \tag{7}$$

Using the shortest path properties,

$$d(v_i, x_c) = \min \left\{ d(v_i, v_p) + d(v_p, x_c), d(v_i, v_q) + d(v_q, x_c) \right\}.$$

We assume vertices are ordered such that:

$$d(v_{i_j}, x_c) = d(v_{i_j}, v_p) + d(v_p, x_c), for\ j = 1, ..., r,$$
$$d(v_{i_j}, x_c) = d(v_{i_j}, v_q) + d(v_q, x_c), for\ j = r + 1, ..., n.$$

Rewriting the objective function and taking derivatives:

$$\begin{aligned}\frac{\partial f}{\partial x_c} =& \sum_{j=1}^{n} h_{i_j} 2d(v_p, x_c) \\ &+ \sum_{j=1}^{r} 2h_{i_j} d(v_{i_j}, v_p) - \sum_{j=r+1}^{n} 2h_{i_j} (d(v_{i_j}, v_q) + d(v_p, v_q)) \qquad (8) \\ \frac{\partial^2 f}{\partial x_c^2} =& \sum_{j=1}^{n} 2h_{i_j}. \qquad (9)\end{aligned}$$

Since (9) is positive, f is convex. Setting (8) to zero and solving for $d(v_p, x_c)$, we get:

$$d(v_p, x_c) = \frac{\sum_{i=r+1}^{n} h_{i_j}(d(v_{i_j}, v_q) + d(v_p, v_q)) - \sum_{i=1}^{r} h_{i_j} d(v_{i_j}, v_p)}{\sum_{i=1}^{n} h_{i_j}}.$$

From this equation, given that the center is to be located on (v_p, v_q), the following interpretations could be made. If $d(v_p, x_c) \leq 0$, x_c will be located on vertex v_p. If $d(v_p, x_c) \geq d(v_p, v_q)$, x_c will be located on vertex v_q. For other values of $d(v_p, x_c)$, x_c will be located on the edge (v_p, v_q). This confirms that the optimal center may lie on an edge rather than at a vertex.

Now we can generalize our results for a network **G**with one cluster. Let us assume that the cluster center is on edge (v_p, v_q) and f_i denotes the objective function component of vertex v_i. There are three cases of f_i as shown in Figure 10. In (a) and (c), the shortest path to the center passes from v_p and v_q, respectively. In (b), when $x_c \in [v_p, b_i]$ where b_i is the arc bottleneck point, the shortest path passes from v_p; otherwise, the shortest path passes from v_q. The function is piecewise and both the function in $(0, b_i)$ and the function in (b_i, l) are convex by second derivative test. Each piece of f_i is convex and increasing with the distance. Because of the convexity of each f_i, f is also convex, which implies that f may contain a local minimum along the edge. Therefore, the optimal solution could be found on the edges.

Theorem 4.1. *In the SSC problem with a single cluster, the optimal cluster center may be located at an interior point of an edge.*

Proof. In order to prove this theorem, we will take objective function value of a vertex which has the best objective function value among vertex set V. Then, we will show that under

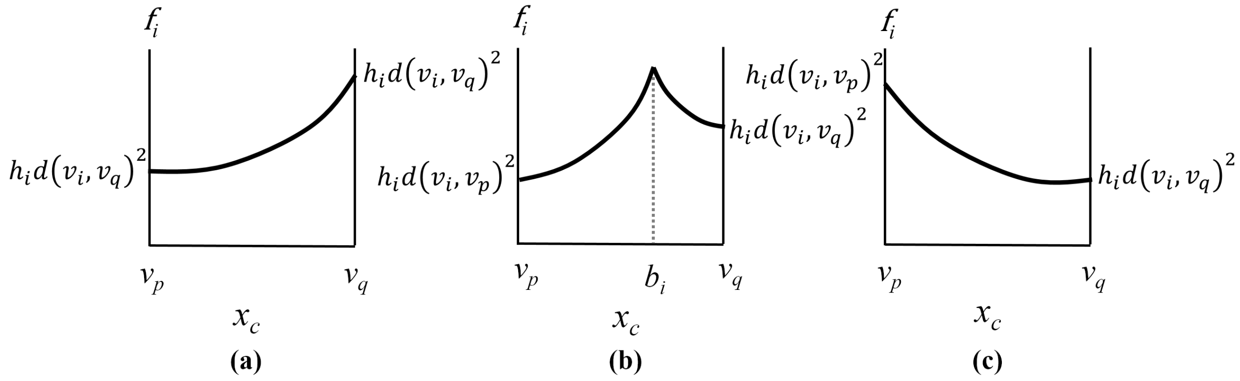

Figure 10: Objective function component for v_i (denoted as f_i) in SSCproblem with single cluster when x_c is moved along the edge (v_p, v_q)

certain conditions, an interior point along the edge will have a lower objective value. Let objective function value at vertex v_p be f^p. Then,

$$f^p = \sum_{i=1}^{n} h_i d(v_i, v_p)^2.$$

Assume that the set of points have been arranged such that for points $j = 1, ..., r$ the shortest path to v_p does not contain the edge (v_p, v_q), and for points $j = r+1, ..., n$, the shortest path to v_p contains the edge (v_p, v_q), that is, $d(v_{i_j}, v_p) = d(v_{i_j}, v_q) + d(v_p, v_q)$. Then,

$$f_p = \sum_{i=1}^{r} h_{i_j} d(v_{i_j}, v_p)^2 + \sum_{i=r+1}^{n} h_{i_j} (d(v_{i_j}, v_q) + d(v_p, v_q))^2.$$

There exists a point x on the edge (v_p, v_q) at which the same partitioning is valid. Then, the objective function value at x is

$$f^x = \sum_{i=1}^{r} h_{i_j} (d(v_{i_j}, v_p) + d(v_p, x))^2 + \sum_{i=r+1}^{n} h_{i_j} (d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, x))^2.$$

Rearranging this expression, we have

$$f^x = f^p + d(v_p, x)^2 \left[\textstyle\sum_{i=1}^{n} h_{i_j}\right] + 2d(v_p, x) \left[\textstyle\sum_{i=1}^{r} h_{i_j} d(v_{i_j}, v_p) - \sum_{i=r+1}^{n} h_{i_j} (d(v_{i_j}, v_q) + d(v_p, v_q))\right].$$

Let

$$a = \sum_{i=1}^{n} h_{i_j}, b = \sum_{i=1}^{r} h_{i_j} d(v_{i_j}, v_p) - \sum_{i=r+1}^{n} h_{i_j}(d(v_{i_j}, v_q) + d(v_p, v_q)).$$

Then, we have

$$f^x = f^p + ad(v_p, x)^2 + 2bd(v_p, x). \tag{10}$$

On the right-hand side, the expression after f^p is a quadratic function of $d(v_p, x)$, that is, $f(x) = a^2 + 2bx$. From (10), we can say that if $f(x) \leq 0$, $z^x <= f^p$. If $f(x) > 0$, $f^x > f^p$. This is possible when $d(v_p, x) \in [0, -2b/a]$. In order for x to have a nonempty interval, $-2b/a \geq 0$. Since a is nonnegative, this is possible if $b \leq 0$. Hence, the following condition must hold.

$$\sum_{i=1}^{r} h_{i_j} d(v_{i_j}, v_p) \leq \sum_{i=r+1}^{n} h_{i_j}(d(v_{i_j}, v_q) + d(v_p, v_q))$$

□

SSC on Networks with p Clusters. When p clusters exist, each vertex is assigned to its closest cluster center, minimizing the sum of squared distances. As cluster centers move along edges, assignments may change at assignment bottleneck points, as illustrated in Figure 11.

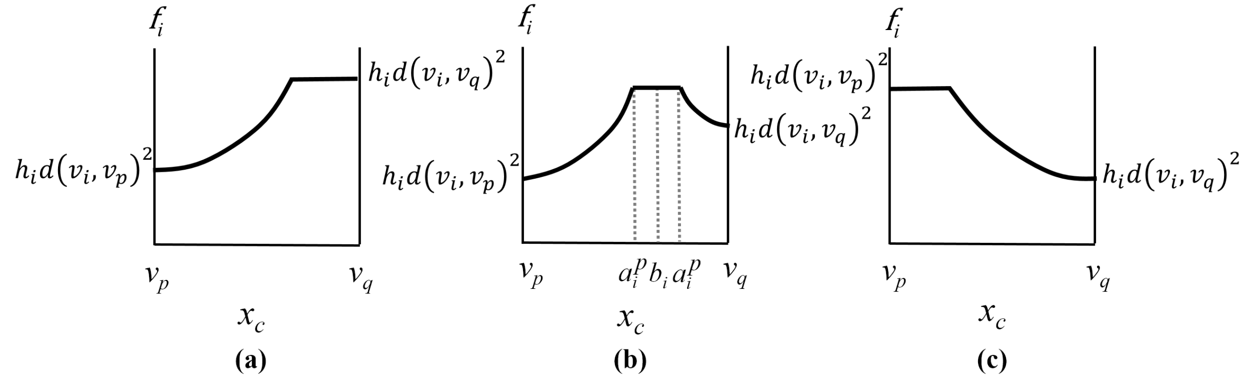

Figure 11: Objective function component for v_i (denoted as f_i) in SSCproblem with p clusters when x_c is moved along the edge (v_p, v_q) and x_k is the second closest cluster center to v_i

Theorem 4.2. *Let* $\mathbf{V}^*$ *be a set of* p *vertices* $\{v_1^*, v_2^*, ..., v_p^*\}$ *which is the optimal solution among all possible* $\mathbf{V}$ *sets. There exists a subset* $\mathbf{X}^* \in \mathbf{G}$ *containing* s *(*$s \leq p$*) centers located* *on edges and the remaining centers at vertices* $\{x_c, x_2, ..., x_s,\}$ $\left\{v^*_{s+1}, v^*_{s+2}, ..., v^*_p\right\}$*, such that*

$$\sum_{k=1}^{n} h_i d(v_i, \mathbf{V}^*)^2 \geq \sum_{k=1}^{n} h_i d(v_i, \mathbf{X})^2$$

Proof. Let $\left\{v^*_1, v^*_2, ..., v^*_p\right\}$ be set of points in $\mathbf{V}^*$. If these points are rearranged such that

$$\begin{aligned} d(v_{i_j}, \mathbf{V}^*) =& d(v_{i_j}, v^*_1) \quad \forall j = 1, ...n_1, \\ d(v_{i_j}, \mathbf{V}^*) =& d(v_{i_j}, v^*_2) \quad \forall j = n_1 + 1, ...n_2, \\ & ... \\ d(v_{i_j}, \mathbf{V}^*) =& d(v_{i_j}, v^*_p) \quad \forall j = n_{p-1} + 1, ...n_p = n, \end{aligned}$$

the objective function could be written as

$$f \quad = \quad \sum_{j=1}^{n_1} h_{i_j} d(v_{i_j}, v^*_1)^2 \quad + \quad \sum_{j=n_1+1}^{n_2} h_{i_j} d(v_{i_j}, v^*_2)^2 \quad + \quad ... \quad + \quad \sum_{j=n_{p-1}+1}^{n_p} h_{i_j} d(v_{i_j}, v^*_p)^2$$

Let

$$\begin{aligned} f_1 =& \sum_{j=1}^{n_1} h_{i_j} d(v_{i_j}, v^*_1)^2 \\ f_2 =& \sum_{j=n_1+1}^{n_2} h_{i_j} d(v_{i_j}, v^*_2)^2 \\ & ... \\ f_p =& \sum_{j=n_{p-1}+1}^{n_p} h_{i_j} d(v_{i_j}, v^*_p)^2. \end{aligned}$$

Define $h'_{i_j} = h'_{i_j}$ for $j = 1, ..., n_1$ and $h'_{i_j} = 0$ for $j = n_1 + 1, ..., n_p$, we have

$$f_1 = \sum_{j=1}^{n} h_{i_j} d(v_{i_j}, v^*_1)^2.$$

In previous theorem, we have shown that given a condition, there exists an interior point x_c

on an edge adjacent to v_1^* such that

$$f_1 \geq \sum_{j=1}^{n} h'_{i_j} d(v_{i_j}, x_c)^2,$$

which could be written as

$$f_1 \geq \sum_{j=1}^{n_1} h_{i_j} d(v_{i_j}, x_c)^2.$$

Assume that for cluster centers 1,...,s, this condition is satisfied. Then,

$$\begin{aligned} f_1 &\geq \sum_{j=1}^{n_1} h_{i_j} d(v_{i_j}, x_c)^2, \\ f_2 &\geq \sum_{j=n_1+1}^{n_2} h_{i_j} d(v_{i_j}, x_2)^2, \\ &... \\ f_s &\geq \sum_{j=n_{s-1}+1}^{n_s} h_{i_j} d(v_{i_j}, x_s)^2. \end{aligned}$$

Adding both sides of the inequalities, we have

$$\begin{aligned} f \geq \sum_{j=1}^{n_1} h_{i_j} d(v_{i_j}, x_c)^2 + &\sum_{j=n_1+1}^{n_2} h_{i_j} d(v_{i_j}, x_2)^2 + ... + \sum_{j=n_{s-1}+1}^{n_s} h_{i_j} d(v_{i_j}, x_s)^2 \\ &+ \sum_{j=n_s+1}^{n_{s+1}} h_{i_j} d(v_{i_j}, v_{s+1}^*)^2 + \sum_{j=n_{s+1}+1}^{n_{s+2}} h_{i_j} d(v_{i_j}, v_{s+2}^*)^2 \\ &+ ... + \sum_{j=n_{p-1}+1}^{n_p} h_{i_j} d(v_{i_j}, v_p^*)^2 \end{aligned} \qquad (11)$$

Let the new set of centers $X = \{x_c, x_2, ..., x_s, v_{s+1}^*, v_{s+2}^*, ..., v_p^*\}$. After changing locations of cluster centers, assignments of vertices to centers may change. Therefore, we have

$$\sum_{j=1}^{n_1} h_{i_j} d(v_{i_j}, x_c)^2 + \sum_{j=n_1+1}^{n_2} h_{i_j} d(v_{i_j}, x_2)^2 + ... + \sum_{j=n_{s-1}+1}^{n_s} h_{i_j} d(v_{i_j}, x_s)^2$$

$$+ \sum_{j=n_s+1}^{n_{s+1}} h_{i_j} d(v_{i_j}, v^*_{s+1})^2 + \sum_{j=n_{s+1}+1}^{n_{s+2}} h_{i_j} d(v_{i_j}, v^*_{s+2})^2$$

$$+ ... + \sum_{j=n_{p-1}+1}^{n_p} h_{i_j} d(v_{i_j}, v^*_p)^2 \geq \sum_{j=1}^{n} h_i d(v_i, \mathbf{X})^2. \tag{12}$$

Combining (11) and (12), we have

$$\sum_{i \in \mathbf{V}} h_i d(v_i, \mathbf{V}^*)^2 \geq \sum_{i \in \mathbf{V}} h_i d(v_i, \mathbf{X})^2.$$

□

This theorem reinforces that restricting cluster centers to vertices may lead to suboptimal solutions in the SSC problem.

4.2. Soft Assignment Problems

In this section, two clustering problems that perform soft assignment will be discussed. In soft assignment, each vertex is assigned to all clusters with a probability. There are two soft clustering problems defined in the scope of this study. Both of these problems have been studied on the plane in the literature, and to the best of our knowledge, they have not been studied on networks before. These two problems are called as Probabilistic Distance Clustering (PD-Clustering) and Fuzzy Clustering (FC). These problems differ in the objective functions and membership functions they use. As a result of this difference, they have different properties, which will be discussed in further detail.

4.2.1. Probabilistic Distance Clustering (PD-Clustering) Problem on Networks

On a network, PD-Clustering Problem is defined as

$$\begin{aligned} &\textit{minimize} \quad f(\mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{p} p_{ik}^2 d(v_i, x_k) && (13) \\ &\quad\text{subject to} \\ &\quad \sum_{k=1}^{p} p_{ik} = 1 \quad \forall\, i = 1, ..., n, \\ &\quad p_{ik} \geq 0 \quad \forall\, i = 1, ..., n,\ k = 1, ..., p, \\ &\quad x_k \in V \quad \forall\, k = 1, ..., p, \\ & && (14) \end{aligned}$$

where x_k is the location of cluster center k and p_{ik} is the membership value of v_i to cluster k. For each vertex, the summation of memberships to all clusters must be equal to 1. As shown in [19] by Iyigun and Ben-Israel, using the Lagrangian Method and keeping all x_k fixed, the optimal membership function is

$$p_{ik}^* = \frac{1}{\sum_{l=1}^{p} \frac{d(v_i,x_k)}{d(v_i,x_l)}}. \qquad (15)$$

Analyzing this problem, we observe that the optimal center locations are on $\mathbf{V}$, which will be proven in this subsection.

PD-Clustering on Networks with a Single Cluster. In PD-Clustering Problem, when there is one cluster, each vertex will be have a membership value of 1 to that cluster. As a result, the problem becomes similar to 1-median problem. f_i, the objective function component of v_i, is as illustrated in Figure 7 in the example illustrated in Figure 6. It is linear and piecewise concave, and its behavior changes at arc bottleneck point if x_c is moved along an edge (v_p, v_q). If there is an arc bottleneck point on an edge as in (b), f_i is linear and piecewise concave along the edge such that it has its maximum at the arc bottleneck point. If there are no arc bottleneck points as in (a) and (c), the distance function is linear. Summation of piecewise concave and linear functions is linear and piecewise concave as well, which is the objective function (13). Therefore, locating the center to an interior point of an edge will lead higher objective function values. The theorem and its proof is given below.

Theorem 4.3. *In single center PD-Clustering Problem,* $\mathbf{V}$*contains the set of optimal solutions.*

Proof. To prove this theorem, it will be shown that an interior point x on an edge (v_p, v_q) could not have a lower objective value than the vertex v_p which has the best objective value among vertex set $\mathbf{V}$. Let objective function value at vertex v_p be f^p. Then,

$$f^p = \sum_{i=1}^{N} p_i^2 d(v_i, v_p).$$

Assume that the set of vertices arranged as the ones whose shortest path to v_p contains the edge (v_p, v_q) or not. Let $v_{i_j}, j = 1, ..., r$ show the vertices that does not contain the edge (v_p, v_q), and $j = r + 1, ..., n$ show the ones that contains the edge (v_p, v_q), that is, $d(v_{i_j}, v_p) = d(v_{i_j}, v_q) + d(v_p, v_q)$. Then,

$$f^p = \sum_{j=1}^{r} p_{i_j}^2 d(v_{i_j}, v_p) + \sum_{j=r+1}^{N} p_{i_j}^2 (d(v_{i_j}, v_q) + d(v_p, v_q)).$$

There exists a center x on the edge (v_p, v_q) at which the same arrangement is valid. Then, the objective function value at x is

$$f^x = \sum_{j=1}^{r} p_{i_j}^2 (d(v_{i_j}, v_p) + d(v_p, x)) + \sum_{j=r+1}^{N} p_{i_j}^2 ((d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, x))).$$

Rearranging this expression, we have

$$f^x = f^p + d(v_p, x) \left[\sum_{j=1}^{r} p_{i_j}^2 - \sum_{j=r+1}^{n} p_{i_j}^2 \right].$$

$$\sum_{j=1}^{r} p_{i_j}^2 \geq \sum_{j=r+1}^{n} p_{i_j}^2 \implies f^p \leq f^x. \tag{16}$$

Suppose that

$$\sum_{j=1}^{r} p_{i_j}^2 < \sum_{j=r+1}^{n} p_{i_j}^2. \tag{17}$$

Since $d(v_p, v_q)$ is a positive constant, multiplying both sides with $d(v_p, v_q)$, we have

$$d(v_p, v_q) \sum_{j=1}^{r} p_{i_j}^2 < d(v_p, v_q) \sum_{j=r+1}^{n} p_{i_j}^2.$$

We may rewrite f^p as

$$f^p = \sum_{j=1}^{r} p_{i_j}^2 d(v_{i_j}, v_p) + \sum_{j=r+1}^{N} p_{i_j}^2 d(v_{i_j}, v_q) + \sum_{j=r+1}^{N} p_{i_j}^2 d(v_p, v_q).$$

By (17), we may write

$$f^p > \sum_{j=1}^{r} p_{i_j}^2 d(v_{i_j}, v_p) + \sum_{j=r+1}^{N} p_{i_j}^2 d(v_{i_j}, v_q) + \sum_{j=1}^{r} p_{i_j}^2 d(v_p, v_q).$$

Right-hand side of this inequality is an upper bound to the objective function value at v_q. Hence,

$$f^p > f^q,$$

which contradicts with v_p's having the minimum objective value among all vertices. This implies that f^x will always be greater than f^p. Therefore, the optimal location will always be on a vertex. □

In the following subsection, a generalized version of this problem, PD-Clustering Problem, with p clusters will be analyzed.

PD-Clustering on Networks with p Clusters. When there are p clusters, PD-Clustering works with membership values which depend on location of centers. In this subsection, it will be proven that in the optimal solution, centers of a PD-Clustering Problem with p clusters on a network will be on vertices.

For the sake of simplicity, suppose we have two clusters. If (15) is evaluated for this case, membership value of vertex i will be

$$p_{i1} = \frac{d(v_i, x_2)}{d(v_i, x_c) + d(v_i, x_2)}, \qquad\qquad p_{i2} = \frac{d(v_i, x_c)}{d(v_i, x_c) + d(v_i, x_2)}. \tag{18}$$

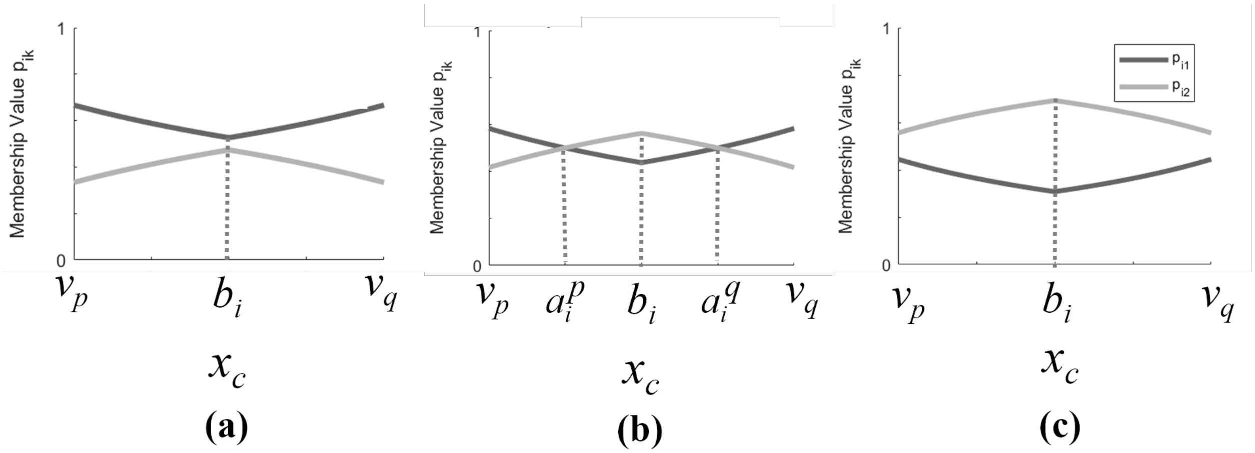

Figure 12: Membership function p_{ik} of PD-Clusteringwith 2 clusters

For a graph $\mathbf{G}$ with two clusters as in Figure 6, keeping x_2 fixed and moving x_c on the edge (v_p, v_q), change in the membership functions p_{i1} and p_{i2} has been visualized in Figure 12. Although x_2 has a fixed location, it is affected from the location change of x_c. In (a), (b) and (c), as x_c moves towards arc bottleneck point b_i, p_{i1} decreases since distance increases and reaches the maximum at b_i. As p_{i1} decreases, p_{i2} increases. In (a), even at the arc bottleneck point, $d(v_i, x_1)$ is less than $d(v_i, x_2)$; therefore, p_{i2} increases but it cannot be greater than p_{i1}. In (c), the contrast occurs. $d(v_i, x_2)$ is less than $d(v_i, x_1)$ even when x_c is located on endpoints v_p or v_q. Therefore, p_{i1} is always less than p_{i2}. In (b), when $x_c \in [v_p, a_i^p]$, p_{i1} is greater than p_{i2}. When $x_c \in [a_i^p, a_i^q]$, p_{i2} is greater than p_{i1} since $d(v_i, x_2)$ is less than $d(v_i, x_1)$. Lastly, as $x_c \in [a_i^q, v_q]$, again, $d(v_i, x_1)$ decreases and becomes less than $d(v_i, x_2)$. Therefore, p_{i1} is greater than p_{i2}. In (b), assignment bottleneck points could be observed as the points where p_{i1}=p_{i2}.

If (18) is substituted in (13), the objective function will be

$$f(\mathbf{X}) = \sum_{i=1}^{n} \frac{d(v_i, x_c) d(v_i, x_2)}{d(v_i, x_c) + d(v_i, x_2)}.$$

For three clusters, the resulting membership values will be

$$\begin{aligned}
p_{i1} &= \frac{d(v_i,x_2)d(v_i,x_3)}{d(v_i,x_c)d(v_i,x_2)+d(v_i,x_2)d(v_i,x_3)+d(v_i,x_c)d(v_i,x_3)},\\
p_{i2} &= \frac{d(v_i,x_c)d(v_i,x_3)}{d(v_i,x_c)d(v_i,x_2)+d(v_i,x_2)d(v_i,x_3)+d(v_i,x_c)d(v_i,x_3)},\\
p_{i3} &= \frac{d(v_i,x_c)d(v_i,x_2)}{d(v_i,x_c)d(v_i,x_2)+d(v_i,x_2)d(v_i,x_3)+d(v_i,x_c)d(v_i,x_3)}.
\end{aligned}$$

With the same manner, the objective function could be written as

$$f(\mathbf{X}) = \sum_{i=1}^{n} \frac{d(v_i,x_c)d(v_i,x_2)d(v_i,x_3)}{d(v_i,x_c)d(v_i,x_2)+d(v_i,x_2)d(v_i,x_3)+d(v_i,x_c)d(v_i,x_3)}.$$

For the problem with p clusters, the objective function is

$$f = \sum_{i=1}^{n} \frac{\prod_{k\in K} d(v_i,x_k)}{\sum_{k\in K}\prod_{l\neq k} d(v_i,x_l)}. \tag{19}$$

Since we assume that location of x_k is fixed for $k = 2, ..., P$, we could separate constant components of each vertex as K_i and write the objective function (19) as in (20).

$$K_i = \frac{\prod_{k=2}^{p} d(v_i,x_k)}{\sum_{k=2}^{p}\prod_{l\neq k} d(v_i,x_l)} \rightarrow \qquad f(x_c) = \sum_{i=1}^{n} \frac{d(v_i,x_c)K_i}{d(v_i,x_c)+K_i} \tag{20}$$

Let the memberships p_{i1} of each point considering the location of cluster center 1 be

$$p_{i1} = \frac{K_i}{d(v_i,x_c)+K_i}. \tag{21}$$

Then, objective function (20) could be simplified as

$$f(x_c) = \sum_{i=1}^{n} d(v_i,x_c)p_{i1}.$$

f_i, contribution of vertex v_i to the function (20), is continuous and twice differentiable. First and second order derivatives are

$$\frac{df_i}{dx_c} = \frac{K_i^2}{(K_i + d(v_i, x_c))^2}$$
$$\frac{d^2 f_i}{dx_c^2} = \frac{-2K_i^2}{(K_i + d(v_i, x_c))^3} \tag{22}$$

With the second derivative test, since (22) is always negative, we can conclude that f_i is concave. Let **G** be a graph with p clusters. Keeping *p-1* clusters fixed and moving one cluster center (let us say x_c) along the edge $(v_p, v - q)$, f_i function could be observed as given in Figure 13. In (a) and (c), the shortest path from v_i to x_c passes from v_p and v_q, respectively. In (b), when $x_c \in [v_p, b_i]$, f_i increases. When $x_c \in [b_i, v_q]$, f_i decreases with the decreasing distance. In (b), f_i is piecewise concave. Different from hard assignment problems, piecewiseness occurs at only arc bottleneck points. Since each f_i is concave or piecewise concave, f, summation of $f_i \, \forall \, i = 1, ..., n$, is also concave.

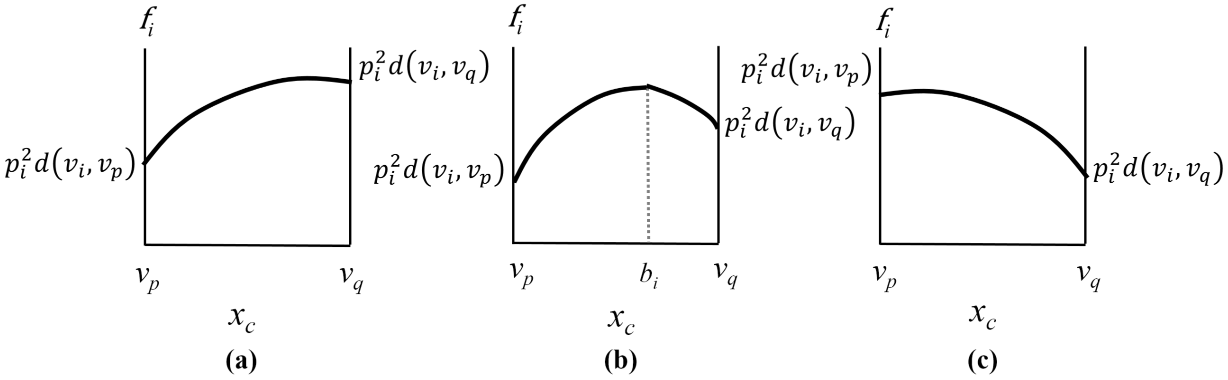

Figure 13: Objective function component for v_i (denoted as f_i) in PD-Clustering Problemwith p clusters when x_c is moved along the edge (v_p, v_q) and x_k is the second closest cluster center to v_i

Theorem 4.4. *In Probabilistic Distance Clustering Problem on Networks, a cluster center is always at a vertex of a network* $\mathbf{G} = (\mathbf{E}, \mathbf{V})$.

Proof. To prove this theorem, it will be shown that keeping $x_k \, \forall \, k = 1, ..., p - \{c\}$ fixed, x_c will always be located on a vertex. Let y_0 be an arbitrary point on the edge $(v_p, v_q) \in \mathbf{E}$and $y_0 \notin V$. There exists a vertex v_m such that

$$\sum_{i=1}^{n} \frac{d(v_i, y_0) K_i}{d(v_i, y_0) + K_i} \geq \sum_{i=1}^{n} \frac{d(v_i, v_m) K_i}{d(v_i, v_m) + K_i}.$$

We know that

$$d(v_i, y_0) = \min\left\{d(v_i, v_p) + d(v_p, y_0), d(v_i, v_q) + d(v_q, y_0)\right\}.$$

Assume that the set of points have been arranged such that

$$\begin{aligned} d(v_{i_j}, y_0) &= d(v_{i_j}, v_p) + d(v_p, y_0), for\ j = 1, ..., r, \\ d(v_{i_j}, y_0) &= d(v_{i_j}, v_q) + d(v_q, y_0), for\ j = r+1, ..., N. \end{aligned}$$

Then, the objective function could be written as

$$\sum_{j=1}^{n} \frac{d(v_{i_j}, y_0)K_{i_j}}{d(v_{i_j}, y_0) + K_{i_j}} = \sum_{j=1}^{r} \frac{(d(v_{i_j}, v_p) + d(v_p, y_0))K_{i_j}}{d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j}} + \sum_{j=r+1}^{n} \frac{(d(v_{i_j}, v_q) + d(v_q, y_0))K_{i_j}}{d(v_{i_j}, v_q) + d(v_q, y_0) + K_{i_j}}. \quad (23)$$

Since we have two vertices v_p and v_q connected by the edge which contains y_0, either v_p or v_q is a better solution. We will consider two cases each implying that one of the vertices is a better solution.

Substitute $d(v_q, y_0) = d(v_p, v_q) - d(v_p, y_0)$ the objective function in (23) is

$$f = \sum_{j=1}^{n} \frac{d(v_{i_j}, y_0)K_{i_j}}{d(v_{i_j}, y_0) + K_{i_j}} = \sum_{j=1}^{r} \frac{(d(v_{i_j}, v_p) + d(v_p, y_0))K_{i_j}}{d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j}} + \sum_{j=r+1}^{n} \frac{(d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0))K_{i_j}}{d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0) + K_{i_j}}.$$

Let

$$f = f_1 + f_2,$$

where

$$f_1 = \sum_{j=1}^{r} \frac{(d(v_{i_j}, v_p) + d(v_p, y_0))K_{i_j}}{d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j}}, \tag{24}$$

$$f_2 = \sum_{j=r+1}^{n} \frac{(d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0))K_{i_j}}{d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0) + K_{i_j}}. \tag{25}$$

Multiply both the numerator and denominator of each term of the summation of f_1 in (24) by $d(v_{i_j}, v_p) + K_{i_j}$, then

$$f_1 = \sum_{j=1}^{r} \left[\frac{d(v_{i_j}, v_p)K_{i_j}}{d(v_{i_j}, v_p) + K_{i_j}} + \frac{d(v_p, y_0)K_{i_j}^2}{(d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_p) + K_{i_j})} \right].$$

Similarly, multiply both the numerator and denominator of each term of the summation of f_2 in (25) by $(d(v_{i_j}, v_q) + d(v_p, v_q) + K_{i_j})$, then

$$f_2 = \sum_{j=r+1}^{n} \left[\frac{(d(v_{i_j}, v_q) + d(v_p, v_q))K_{i_j}}{d(v_{i_j}, v_q) + d(v_p, v_q) + K_{i_j}} - \frac{d(v_p, y_0)K_{i_j}^2}{(d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_q) + d(v_p, v_q) + K_{i_j})} \right]. \tag{26}$$

By triangle inequality, we have $d(v_{i_j}, v_q) + d(v_p, v_q) \geq d(v_{i_j}, v_p)$. Then

$$\sum_{j=r+1}^{n} \frac{(d(v_{i_j}, v_q) + d(v_p, v_q))K_{i_j}}{d(v_{i_j}, v_q) + d(v_p, v_q) + K_{i_j}} \geq \sum_{j=r+1}^{n} \frac{d(v_{i_j}, v_p)K_{i_j}}{d(v_{i_j}, v_p) + K_{i_j}}. \tag{27}$$

Substitute right-hand side of (27) in (26), we have

$$f_2 \geq f_2' = \sum_{j=r+1}^{n} \left[\frac{d(v_{i_j}, v_p)K_{i_j}}{d(v_{i_j}, v_p) + K_{i_j}} - \frac{d(v_p, y_0)K_{i_j}^2}{(d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_q) + d(v_p, v_q) + K_{i_j})} \right]. \tag{28}$$

Since $d(v_{i_j}, v_q) + d(v_p, v_q) \geq d(v_{i_j}, v_q)$, replace $d(v_{i_j}, v_q) + d(v_p, v_q)$ in (28) with $d(v_{i_j}, v_q)$, then

$$f'_2 \geq f''_2 = \sum_{j=r+1}^{n} \left[\frac{d(v_{i_j}, v_p) K_{i_j}}{d(v_{i_j}, v_p) + K_{i_j}} - \frac{d(v_p, y_0) K_{i_j}^2}{(d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_q) + K_{i_j})} \right].$$

Hence,

$$f \geq f_1 + f''_2 \tag{29}$$

If (29) is rearranged, then

$$f \geq \sum_{j=1}^{n} \frac{d(v_{i_j}, v_p) K_{i_j}}{d(v_{i_j}, v_p) + K_{i_j}} \tag{30}$$

$$+ d(v_p, y_0) \left[\sum_{j=1}^{r} \frac{K_{i_j}^2}{(d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_p) + K_{i_j})} \right. \tag{31}$$

$$\left. - \sum_{j=r+1}^{n} \frac{K_{i_j}^2}{(d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_q) + K_{i_j})} \right]. \tag{32}$$

Summations in (31) and (32) are equal to (33) and (34), respectively.

$$\sum_{j=1}^{r} p_{iy_0} p_{iv_p} \tag{33}$$

$$\sum_{j=r+1}^{n} p_{iy_0} p_{iv_q} \tag{34}$$

If (35) is satisfied, (36) will hold true. That is, v_p is a better location than y_0 for x_c.

$$\sum_{j=1}^{r} p_{iy_0} p_{iv_p} \geq \sum_{j=r+1}^{n} p_{iy_0} p_{iv_q} \tag{35}$$

$$\sum_{j=1}^{n} \frac{d(v_{i_j}, y_0) K_{i_j}}{d(v_{i_j}, y_0) + K_{i_j}} \geq \sum_{j=1}^{n} \frac{d(v_{i_j}, v_p) K_{i_j}}{d(v_{i_j}, v_p) + K_{i_j}} \tag{36}$$

If (35) is not satisfied, then we have

$$\sum_{j=1}^{r} p_{iy_0} p_{iv_p} < \sum_{j=r+1}^{n} p_{iy_0} p_{iv_q} \tag{37}$$

Clearly, (36) is not guaranteed in this case.

Again, let

$$f = f_1 + f_2$$

Add and subtract $d(v_p, v_q)$ both numerators and denominators of each term of summation of f_1 in (24), then

$$f_1 = \sum_{j=1}^{r} \frac{(d(v_{i_j}, v_p) + d(v_p, y_0) + d(v_p, v_q) - d(v_p, v_q))K_{i_j}}{d(v_{i_j}, v_p) + d(v_p, y_0) + d(v_p, v_q) - d(v_p, v_q) + K_{i_j}}, \tag{38}$$

Multiply both the numerator and denominator of each term of the summation of f_1 in (38) by $d(v_{i_j}, v_p) + d(v_p, v_q) + K_{i_j}$, then

$$\begin{aligned} f_1 = \sum_{j=1}^{r} \Bigg[& \frac{(d(v_{i_j}, v_p) + d(v_p, v_q))K_{i_j}}{d(v_{i_j}, v_p) + d(v_p, v_q) + K_{i_j}} \\ & - \frac{(d(v_p, v_q) - d(v_p, y_0))K_{i_j}^2}{(d(v_{i_j}, v_p) + d(v_p, v_q) - d(v_p, v_q) + d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_p) + d(v_p, v_q) + K_{i_j})} \Bigg]. \end{aligned}$$

Cancelling the $d(v_p, v_q) - d(v_p, v_q)$ expression, we have

$$f_1 = \sum_{j=1}^{r} \Bigg[\frac{(d(v_{i_j}, v_p) + d(v_p, v_q))K_{i_j}}{d(v_{i_j}, v_p) + d(v_p, v_q) + K_{i_j}} \tag{39}$$

$$- \frac{(d(v_p, v_q) - d(v_p, y_0))K_{i_j}^2}{(d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_p) + d(v_p, v_q) + K_{i_j})} \Bigg]. \tag{40}$$

Similarly, multiply both the numerator and denominator of each term of the summation of

f_2 in (25) by $(d(v_{i_j}, v_q) + K_{i_j})$, then

$$f_2 = \sum_{j=r+1}^{n} \Bigg[\frac{d(v_{i_j}, v_q)K_{i_j}}{d(v_{i_j}, v_q) + K_{i_j}} + \frac{((d(v_p, v_q) - d(v_p, y_0))K_{i_j}^2}{(d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_q) + K_{i_j})} \Bigg].$$

By triangle inequality, we have $d(v_{i_j}, v_p) + d(v_p, v_q) \geq d(v_{i_j}, v_q)$. Then

$$\sum_{j=1}^{r} \frac{(d(v_{i_j}, v_p) + d(v_p, v_q))K_{i_j}}{d(v_{i_j}, v_p) + d(v_p, v_q) + K_{i_j}} \geq \sum_{j=1}^{r} \frac{d(v_{i_j}, v_q)K_{i_j}}{d(v_{i_j}, v_q) + K_{i_j}}. \tag{41}$$

Substitute right-hand side of (41) in (39), we have

$$f_1 \geq f_1' = \sum_{j=1}^{r} \Bigg[\frac{d(v_{i_j}, v_q)K_{i_j}}{d(v_{i_j}, v_q) + K_{i_j}} - \frac{(d(v_p, v_q) - d(v_p, y_0))K_{i_j}^2}{(d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_p) + d(v_p, v_q) + K_{i_j})} \Bigg]. \tag{42}$$

Since $d(v_{i_j}, v_p) + d(v_p, v_q) \geq d(v_{i_j}, v_p)$, replace $d(v_{i_j}, v_p) + d(v_p, v_q)$ in (42) with $d(v_{i_j}, v_p)$, then

$$f_1' \geq f_1'' = \sum_{j=1}^{r} \Bigg[\frac{d(v_{i_j}, v_q)K_{i_j}}{d(v_{i_j}, v_q) + K_{i_j}} - \frac{(d(v_p, v_q) - d(v_p, y_0))K_{i_j}^2}{(d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_p) + K_{i_j})} \Bigg].$$

Hence,

$$z \geq f_1'' + f_2 \tag{43}$$

If (43) is rearranged, then

$$\begin{aligned}
z \geq & \sum_{j=1}^{n} \frac{d(v_{i_j}, v_q)K_{i_j}}{d(v_{i_j}, v_q) + K_{i_j}} \\
& + (d(v_p, v_q) - d(v_p, y_0)) && (44)\\
& * \left[\sum_{j=r+1}^{n} \frac{K_{i_j}^2}{(d(v_{i_j}, v_q) + d(v_p, v_q) - d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_q) + K_{i_j})} \right. && (45)\\
& \left. - \sum_{j=1}^{r} \frac{K_{i_j}^2}{(d(v_{i_j}, v_p) + d(v_p, y_0) + K_{i_j})(d(v_{i_j}, v_p) + K_{i_j})} \right]. && (46)
\end{aligned}$$

Summations in (45) and (46) are equal to (33) and (34), respectively. If (37) is satisfied, (47) will hold true. v_q is a better location than y_0 for x_c.

$$\sum_{i=1}^{n} \frac{d(v_i, y_0)K_i}{d(v_{i_j}, y_0) + K_i} \geq \sum_{i=1}^{n} \frac{d(v_i, v_q)K_i}{d(v_{i_j}, v_q) + K_i} \tag{47}$$

As shown above, when (35) is satisfied. v_p is a better location than y_0 for x_c. Otherwise, (when (37) is satisfied), v_q is a better location than y_0 for x_c. As a result, the center x_c will always be located on a vertex. □

This proof supports Levy's proof that in the case of concavity, the center will be located on $\mathbf{V}$. This result could be generalized such that all the cluster centers are located on vertices in optimal solution.

Theorem 4.5. *For every cluster center* $\{x_c, x_2, ..., x_p\} \in G$*, there exists* $\{v_{m_1}, v_{m_2}, ..., v_{m_p}\} \in V$ *such that*

$$\sum_{k=1}^{P} \sum_{i=1}^{n} p_{ik}^2 d(v_i, x_k) \geq \sum_{k=1}^{P} \sum_{i=1}^{n} p_{ik}^2 d(v_i, v_k)$$

Proof. To prove this theorem, we will change location of a center i (x_i) by fixing the other centers ($x_k, k \neq i$). In each location change, membership scores will change. Initial membership score will be denoted as $p_{ik}^{(0)}$. Updated membership scores resulting from changing location of cluster center k will be denoted as $p_{ik}^{(k)}$.

Let x_c be the cluster center to be changed, keeping the others fixed. In previous theorem, we have shown that

$$\sum_{k=1}^{P}\sum_{i=1}^{n} p_{ik}^{(0)^2} d(v_i, x_k) \quad \geq \quad \sum_{i=1}^{n} p_{ik}^{(1)^2} d(v_i, v_{m_1}) \quad + \quad \sum_{k=2}^{P}\sum_{i=1}^{n} p_{ik}^{(1)^2} d(v_i, x_k). \quad (48)$$

Now, let x_2 be the cluster center to be changed, keeping the others fixed in their last locations. Again, we have

$$\sum_{i=1}^{n} p_{ik}^{(1)^2} d(v_i, v_{m_1}) + \sum_{k=2}^{P}\sum_{i=1}^{n} p_{ik}^{(1)^2} d(v_i, x_k) \geq$$

$$\sum_{k=1}^{2}\sum_{i=1}^{n} p_{ik}^{(2)^2} d(v_i, v_{m_k}) + \sum_{k=3}^{P}\sum_{i=1}^{n} p_{ik}^{(2)^2} d(v_i, x_k). \quad (49)$$

Perform this process with each x_k as the center location to be changed, as the last expression, we have

$$\sum_{k=1}^{p-1}\sum_{i=1}^{n} p_{ik}^{(p-1)^2} d(v_i, v_{m_k}) \quad + \quad \sum_{i=1}^{n} p_{ik}^{(p-1)^2} d(v_i, x_p) \quad \geq \quad \sum_{k=1}^{p}\sum_{i=1}^{n} p_{ik}^{(p)^2} d(v_i, v_{m_k}). \quad (50)$$

Combine (48), (49) and (50), we have

$$\sum_{k=1}^{P}\sum_{i=1}^{n} p_{ik}^{(0)^2} d(v_i, x_k) \geq \sum_{k=1}^{P}\sum_{i=1}^{n} p_{ik}^{(p)^2} d(v_i, v_{m_k}),$$

which implies that as center locations, $v_{m_1}, ..., v_{m_k}$ lead to a better solution than $x_c, ..., x_k$. □

As a result, it has been shown that in PD-Clustering Problemon Networks, the optimal solution will always be located on vertices. Therefore, one would not lose from objective function if they restrict center locations as **V** instead of **G**.

4.2.2. Fuzzy Clustering (FC) Problem on Networks

Fuzzy Clustering Problem on Networks is defined as

$$minimize \quad f(\mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{p} p_{ik}^{m} d(v_i, x_k)^2 \tag{51}$$

$$\text{subject to}$$

$$\sum_{k=1}^{p} p_{ik} = 1 \quad \forall\, i = 1, ..., n,$$

$$p_{ik} \geq 0 \quad \forall\, i = 1, ..., n,\ k = 1, ..., p,$$

$$x_k \in G \quad \forall\, k = 1, ..., p,$$

where x_k is the location of cluster center k and p_{ik} is the membership value of v_i to cluster k. m is called *fuzzifier*, or *fuzziness index*. It determines the level of fuzziness in memberships. If $m = 1$, the problem becomes a hard assignment problem - to be more precise, SSCproblem. As m gets larger, all membership values converge to $1/p$. For each vertex, summation of memberships to all clusters must be equal to 1. Derived by Bezdek et.al. in [10] with the use of Lagrangian, keeping all x_k fixed, membership function is

$$p_{ik}^{*} = \frac{1}{\sum_{l=1}^{p}\left(\frac{d(v_i,x_k)}{d(v_i,x_l)}\right)^{\frac{2}{(m-1)}}}. \tag{52}$$

When this problem has been investigated, it has been observed that the optimal center locations are could be on anywhere on the $\mathbf{G}$. In this subsection, this property will be analyzed.

Fuzzy Clustering on Networks with a Single Cluster. As in PD-Clustering if there is a single cluster, all vertices will have a membership equal to 1. FC with 1 cluster differs from PD-Clusteringin that (51), becomes sum of squared distances. Therefore, the problem will demonstrate characteristics of SSC problem with 1 cluster. In a $\mathbf{G}$ with one cluster x_c moving along the edge (v_p, v_q), f_i, the objective function component of v_i, will be as in Figure 10. f_i is a second degree polynomial function increasing with $d(v_i, x_k)$. f_i is convex or piecewise on an edge. This piecewiseness occur at arc bottleneck points. Each piece of f_i is convex and increasing with distance. Because of the convexity, f which is summation of f_i

functions is also convex. But it is not monotone; therefore, it may contain a local minimum along an edge. As a result, there could be an optimal center location located on interior point of an edge. Based on this observation, we can conclude that the following theorem holds.

Theorem 4.6. *Let* $\mathbf{V}^*$ *be a set of* p *vertices* $\left\{v_1^*, v_2^*, ..., v_p^*\right\}$ *which is the optimal solution among all possible* $\mathbf{V}$*sets. In Fuzzy Clustering Problem on networks with a single cluster, there exists a subset* $\mathbf{X}^* \in \mathbf{G}$ *containing centers located on edges such that it has an objective function value lower than* $\mathbf{V}^*$.

Fuzzy Clustering on Networks with p Clusters. In this subsection, objective function of FC Problem with p clusters will be analyzed and structural properties will be investigated.

For the sake of simplicity, suppose we have two clusters. If (52) is evaluated for this case, membership value of vertex i will be

$$p_{i1} = \frac{d(v_i, x_2)^{\frac{2}{(m-1)}}}{d(v_i, x_c)^{\frac{2}{(m-1)}} + d(v_i, x_2)^{\frac{2}{(m-1)}}}, \qquad p_{i2} = \frac{d(v_i, x_c)^{\frac{2}{(m-1)}}}{d(v_i, x_c)^{\frac{2}{(m-1)}} + d(v_i, x_2)^{\frac{2}{(m-1)}}}. \tag{53}$$

For a graph $\mathbf{G}$ with two clusters as in Figure 6, keeping x_2 fixed and moving x_c on the edge (v_p, v_q), change in the membership functions p_{i1} and p_{i2} has been visualized in Figure 14 with fuzziness index m value of 2. As in PD-Clustering, both memberships are affected by the location change of x_c. In (a), (b) and (c), as x_c moves towards arc bottleneck point b_i, p_{i1} decreases due to the increase in distance. As p_{i1} decreases, p_{i2} increases. (a) illustrates the case $d(v_i, x_1)$ is less than $d(v_i, x_2)$; therefore, p_{i2} is less than p_{i1}. (c) is the case $d(v_i, x_1)$ is less than $d(v_i, x_2)$; as a result, p_{i1} is less than p_{i2}. In (b), if $x_c \in [v_p, a_i^p]$, p_{i1} is greater than p_{i2}. If $x_c \in [a_i^p, a_i^q]$, p_{i2} is greater than p_{i1}. In the last interval which is $x_c \in [a_i^q, v_q]$, $d(v_i, x_1)$ is less than $d(v_i, x_2)$. Hence, p_{i1} is greater than p_{i2}. In (b), assignment bottleneck points could be observed as the points where p_{i1}=p_{i2}. Figure 15 is drawn with fuzziness index $m = 20$ to illustrate effect of increase in m in membership function. As could be observed, it does not change the behavior of the membership function. However, even at the points where $d(v_i, x_1)$ values have the maximum difference, memberships p_{i1} and p_{i2} are very close to each other compared to the case of m=2. The effect of increase in m is increase in the fuzziness of the memberships.

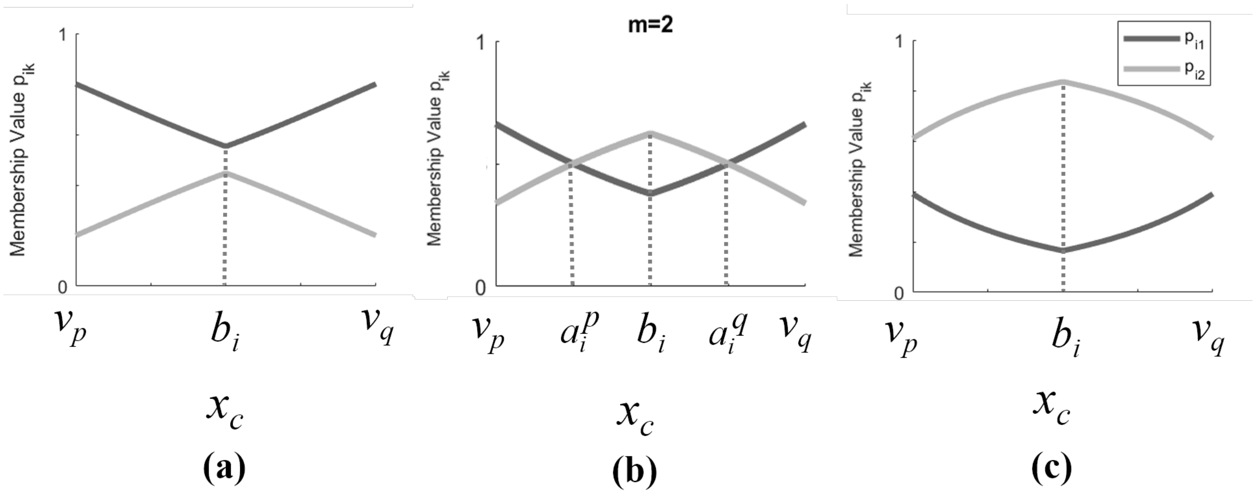

Figure 14: Membership function p_{ik} of FC with 2 clusters when m=2

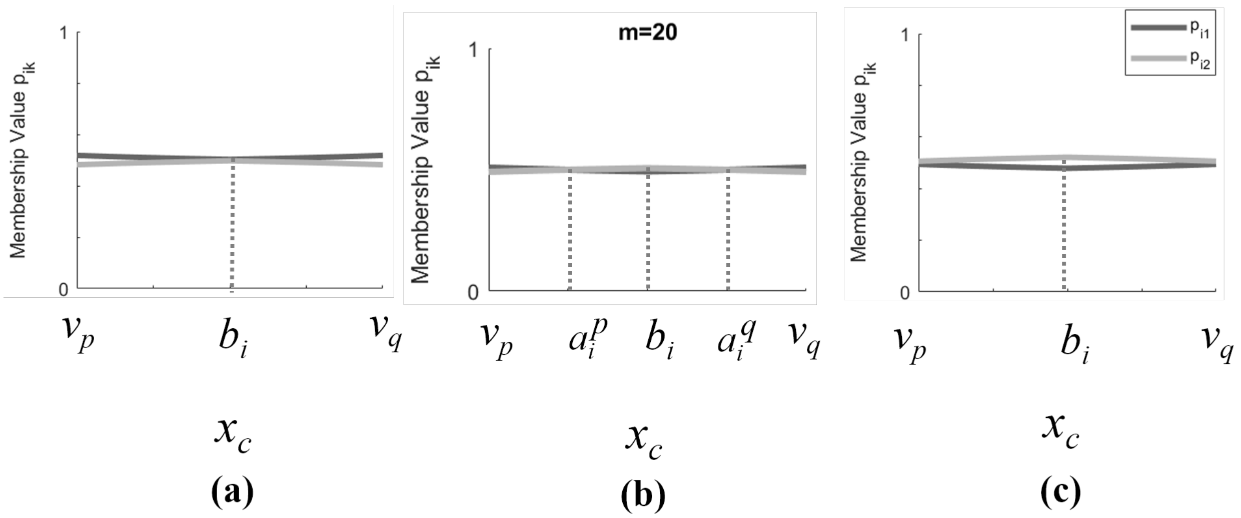

Figure 15: Membership function p_{ik} of FC with 2 clusters when m=20

If (18) is substituted in (51), the objective function will be

$$f(\mathbf{X}) = \sum_{i=1}^{n} \frac{d(v_i, x_c)^2 d(v_i, x_2)^2}{\left(d(v_i, x_c)^{\frac{2}{(m-1)}} + d(v_i, x_2)^{\frac{2}{(m-1)}}\right)^{m-1}}.$$

For three clusters, the membership values are

$$p_{i1} = \frac{(d(v_i,x_2)d(v_i,x_3))^{\frac{2}{(m-1)}}}{(d(v_i,x_c)d(v_i,x_2))^{\frac{2}{(m-1)}} + (d(v_i,x_2)d(v_i,x_3))^{\frac{2}{(m-1)}} + (d(v_i,x_c)d(v_i,x_3))^{\frac{2}{(m-1)}}},$$
$$p_{i2} = \frac{(d(v_i,x_c)d(v_i,x_3))^{\frac{2}{(m-1)}}}{(d(v_i,x_c)d(v_i,x_2))^{\frac{2}{(m-1)}} + (d(v_i,x_2)d(v_i,x_3))^{\frac{2}{(m-1)}} + (d(v_i,x_c)d(v_i,x_3))^{\frac{2}{(m-1)}}},$$
$$p_{i3} = \frac{(d(v_i,x_c)d(v_i,x_2))^{\frac{2}{(m-1)}}}{(d(v_i,x_c)d(v_i,x_2))^{\frac{2}{(m-1)}} + (d(v_i,x_2)d(v_i,x_3))^{\frac{2}{(m-1)}} + (d(v_i,x_c)d(v_i,x_3))^{\frac{2}{(m-1)}}}.$$

With the same manner, the objective function could be written as

$$f(\mathbf{X}) = \sum_{i=1}^{n} \frac{(d(v_i,x_c)d(v_i,x_2)d(v_i,x_3))^2}{\left((d(v_i,x_c)d(v_i,x_2))^{\frac{2}{(m-1)}} + (d(v_i,x_2)d(v_i,x_3))^{\frac{2}{(m-1)}} + (d(v_i,x_c)d(v_i,x_3))^{\frac{2}{(m-1)}}\right)^{m-1}}.$$

In the version of the problem with p clusters, the objective function is

$$f = \sum_{i=1}^{n} \frac{\prod_{k\in K} d(v_i,x_k)^2}{\left(\sum_{k\in K}\prod_{l\neq k} d(v_i,x_l)^{\frac{2}{(m-1)}}\right)^{m-1}} \tag{54}$$

For the sake of simplicity, we assume that $m = 2$. Since we assume that location of x_k is fixed for $k = 1, ..., p - \{c\}$, we could separate constant components of each vertex as K_i and write the objective function (54) as in (55).

$$K_i = \frac{\prod_{k=2}^{p} d(v_i,x_k)^2}{\sum_{k=2}^{p}\prod_{l\neq k} d(v_i,x_l)^2} \rightarrow \qquad f(x_c) = \sum_{i=1}^{n} \frac{d(v_i,x_c)^2 K_i}{d(v_i,x_c)^2 + K_i} \tag{55}$$

f_i, contribution of vertex v_i to the function (20), is continuous and twice differentiable. First and second order derivatives are

$$\begin{aligned} \frac{df_i}{dx_c} &= \frac{2K_i^2}{(K_i + d(v_i,x_c)^2)^2} \\ \frac{d^2 f_i}{dx_c^2} &= \frac{(2K_i^2)(K_i - 3d(v_i,x_c)^2)}{(K_i + d(v_i,x_c)^2)^3}. \end{aligned} \tag{56}$$

With the second derivative test, f_i is

- Convex if $\frac{d^2 f_i}{dx_c^2} \geq 0$, that is, $d(v_i, x_c) \leq \sqrt{\frac{K_i}{3}}$,
- Concave if $\frac{d^2 f_i}{dx_c^2} \geq 0$, that is, $d(v_i, x_c) \geq \sqrt{\frac{K_i}{3}}$.

Theorem 4.7. *In Fuzzy Clustering Problemwith p clusters, given a solution* $\mathbf{X}= \{x_1, x_2, ..., x_p\}$*, if* $x_c \in \mathbf{X}$ *is moved along a given edge keeping the centers* $\mathbf{X}{-}x_c$ *fixed, there could be a location on the given edge that minimizes the objective function* (54)*.*

Proof. Let

$$c_i = \prod_{k=2}^{p} d(v_i, x_k)$$
$$t_i = \sum_{k=2}^{p} \prod_{j \in K, j \neq k} d(v_i, x_j)^{\frac{2}{m-1}}.$$

If we fix locations of x_k, $k \in K - \{1\}$ and separate fixed components of f_i from variable components by using c_i and t_i, the objective function is (57).

$$f_i = \frac{c_i^2 d(v_i, x_c)^2}{\left(x^{\frac{2}{m-1}} t_i + c_i^{\frac{2}{m-1}}\right)^{m-1}} \tag{57}$$

If we calculate first and second order derivatives for f_i with any $m > 1$, we have

$$\frac{df_i}{dx_c} = \frac{(2c_i^2 d(v_i, x_c))}{(c_i + t_i d(v_i, x_c)^{\frac{2}{m-1}})^m} \tag{58}$$

$$\frac{d^2 f_i}{dx_c^2} = \frac{(2c_i^2)\left(c_i(m-1) - t_i(m+1) d(v_i, x_c)^{\frac{2}{m-1}}\right)}{\left(\left(c_i + t_i d(v_i, x_c)^{\frac{2}{m-1}}\right)(m-1)\right)^{m+1}} \tag{59}$$

By the second derivative test, f_i is

- Convex if $\frac{d^2 f_i}{dx_c^2} \geq 0$, that is, $d(v_i, x_c) \geq \sqrt[\frac{2}{m-1}]{\frac{c_i(m-1)}{t_i(m+1)}}$,
- Concave if $\frac{d^2 f_i}{dx_c^2} \geq 0$, that is, $d(v_i, x_c) < \sqrt[\frac{2}{m-1}]{\frac{c_i(m-1)}{t_i(m+1)}}$.

As a result, f_i is a nonconvex function of $d(v_i, x_k)$, and increasing with distance. The objective function could be illustrated as in Figure 16. As in PD-Clustering, f_i is piecewise at arc bottleneck points. And each piece of f_i is nonconvex according to the sign of second derivative (59). Since summation of nonconvex functions are nonconvex, f, summation of $f_i \, \forall \, i = 1, ..., n$, is nonconvex. Since f is not monotone, local minimum could be found at a point where second derivative is positive and first derivative is zero.

Let **G** be a graph with p clusters. Keeping *p-1* clusters fixed and moving one cluster center (let us say x_c) along the edge (v_p, v_q), f_i function could be observed as given in Figure 13. In (a) and (c), the shortest path from v_i to x_c passes from v_p and v_q, respectively. In (b), when $x_c \in [v_p, b_i]$, f_i increases. When $x_c \in [b_i, v_q]$, f_i decreases with the decreasing distance. In (b), f_i is piecewise concave. Different from hard assignment problems, piecewiseness occurs at only arc bottleneck points only. Since each f_i is concave or piecewise concave, f, summation of $f_i \, \forall \, i = 1, ..., n$, is also concave.

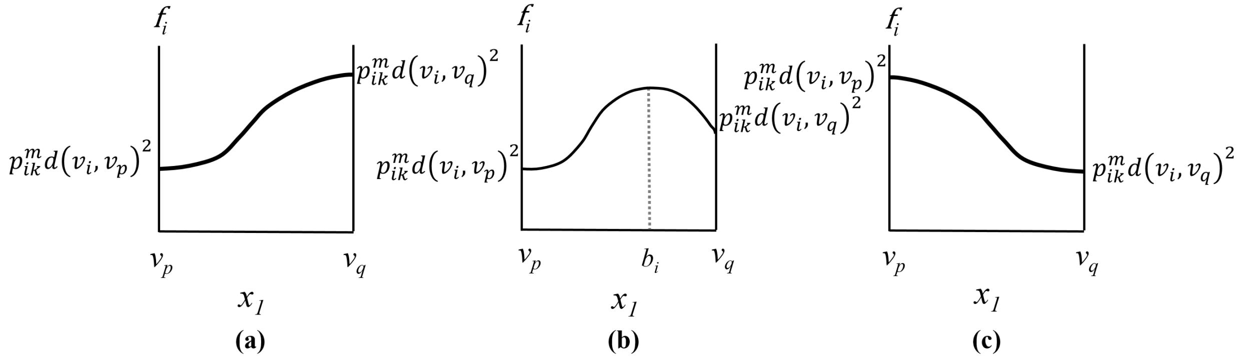

Figure 16: Objective function component for v_i (denoted as f_i) in FC problem with p clusters when x_c is moved along the edge (v_p, v_q) and x_k is the second closest cluster center to v_i

Hence, in Fuzzy Clustering Problem one may find an optimal solution that contains cluster centers located at edges.

□

5. Discussion

The theoretical findings presented in this study offer a fresh perspective on clustering and network location problems by revealing structural properties that influence optimal solutions.

The results highlight the impact of assignment strategies—whether hard or soft—on where cluster centers should be located within a network. These findings challenge conventional assumptions, particularly the idea that optimal clustering solutions can be freely placed on the network space. Instead, we have shown that in certain cases, such as in the PD-Clustering problem, the optimal solution is always restricted to network vertices, while in others, such as the SSC problem, optimal locations may exist on the edges as well.

A key insight of this study is the introduction of assignment bottleneck points, which dictate how assignments shift as cluster centers move along network edges. This property fundamentally changes how clustering solutions should be approached, as it determines regions where abrupt assignment changes occur. Additionally, the verification of vertex optimality in soft assignment clustering problems provides theoretical justification for restricting solution searches to discrete network points rather than treating the network as a continuous space. These properties not only refine our understanding of network clustering but also suggest efficient ways to structure solution algorithms.

From an optimization standpoint, these findings open up new avenues for designing more effective heuristics and metaheuristics. The fact that certain clustering problems always yield optimal solutions at vertices implies that computational effort can be significantly reduced by focusing searches on discrete locations rather than the entire network. For problems where optimal solutions may exist on edges, incorporating assignment bottleneck points into search strategies can lead to improved heuristics that avoid exhaustive enumeration while still capturing high-quality solutions. These theoretical insights provide a foundation for developing practical algorithms that balance efficiency and accuracy, especially in large-scale networks.

The implications of these findings extend beyond theoretical optimization, with direct applicability to real-world problems where clustering decisions must be made on constrained networks. In urban planning, these results can inform optimal placement strategies for public service facilities, such as fire stations or hospitals, by ensuring they are located in positions that minimize response times. In telecommunications and sensor networks, where signals must be optimally distributed across a predefined infrastructure, understanding whether solutions should be placed at vertices or along edges can improve network efficiency. Similarly, in supply chain logistics, warehouse and distribution center locations can be determined more effectively by incorporating assignment dynamics into network-based clustering models. The ability to leverage these theoretical properties in real-world settings has the potential to improve decision-making processes in various domains where spatial optimization is critical.

6. Conclusion

This study establishes fundamental theoretical properties that govern clustering problems on networks, leading to a more structured understanding of optimal center placement. The identification of assignment bottleneck points and the distinction between vertex-restricted and edge-admissible solutions provide a new perspective on how clustering problems should be formulated and solved. By demonstrating that probabilistic distance clustering (PD-Clustering) problems always have optimal solutions at vertices, we provide theoretical justification for discrete optimization approaches in these settings. Conversely, by showing that sum of squares clustering (SSC) solutions may lie on edges, we emphasize the importance of considering a broader search space in these cases.

These findings suggest that traditional plane-based clustering methods, which assume free placement of centers, cannot always be directly applied to networks. Instead, clustering problems in network-constrained environments require careful consideration of assignment dynamics and structural properties. The results of this study can guide the development of more efficient solution approaches, including heuristic and metaheuristic methods that take advantage of these theoretical insights.

Future research should explore algorithmic techniques that leverage the properties identified in this study to develop practical, scalable solutions. Given the broad applicability of clustering problems across transportation, telecommunications, emergency response, and logistics, understanding the fundamental structure of optimal solutions can lead to better decision-making in real-world scenarios. By building on these theoretical results, future studies can bridge the gap between mathematical optimization and practical implementation, ensuring that network-based clustering problems are addressed with both precision and computational efficiency.

References

[1] J. A. Hartigan, Clustering Algorithms, John Wiley and Sons, 1975.

[2] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: a review, ACM computing surveys (CSUR) 31 (3) (1999) 264–323.

[3] S. Ayramo, T. Karkkainen, Introduction to partitioning based clustering methods with a robust example, Report (2006).

[4] L. Ou-Yang, H. Yan, X.-F. Zhang, A multi-network clustering method for detecting protein complexes from multiple heterogeneous networks, BMC Bioinformatics (2017). doi:10.1186/s12859-017-1877-4.

[5] S. E. Schaeffer, Graph clustering, Computer Science Review I (2007) 27–64.

[6] R. Tariq, K. Lavangnananda, P. Bouvry, P. Mongkolnam, Partitioning graph clustering with user-specified density, IEEE Access 11 (2023) 122273–122274. doi:10.1109/ACCESS.2023.3329429.

[7] M. Zaferanieh, M. Abareshi, A traffic-based model to the p-median problem in congested networks, Communications in Combinatorics and Optimization xx (x) (2024) 1–27. doi:10.22049/cco.2024.29105.1992.

[8] Y. Mahmoudi, N. Zioui, H. Belbachir, H. Dagdougui, B. Said, Reducing wireless sensors networks energy consumption using p-median modelling and optimization, Operational Research in Engineering Sciences: Theory and Applications 6 (2) (2023) 180–196. doi:10.31181/oresta/060210.

[9] R. Hurtado, J. Bobadilla, R. Bojorque, F. Ortega, X. Li, A new recommendation approach based on probabilistic soft clustering methods: A scientific documentation case study, IEEE Access 7 (2019) 7522–7536. doi:10.1109/ACCESS.2018.2890079.

[10] J. C. Bezdek, R. Ehrlich, W. Full, Fcm: The fuzzy c-means clustering algorithm, Computers and Geosciences 10 (2-3) (1984) 191–203.

[11] H. J. Zimmermann, Fuzzy set theory, WIREs Computational Statistics 2 (2010).

[12] C. Iyigun, A. Ben-Israel, Probabilistic d-clustering, Journal of Classification 25 (2008) 5–26.

[13] C. C. Aggarwal, H. Wang, Managing and Mining Graph Data, Springer Science+Business Media, 2010, Ch. A Survey of Clustering Algorithms for Graph Data.

[14] S. Hakimi, Optimum locations of switching centers and the absolute centers and medians of a graph, Operations Research 12 (3) (1964) 450–459.

[15] S. L. Hakimi, Optimum distribution of switching centers in a communication network and some related graph theoretic problems, Operational Research 13 (1965) 462–475.

[16] J. Levy, An extended theorem for location on a network, Operational Research Quarterly 18 (4) (1967) 433–442.

[17] J. N. Hooker, R. S. Garfinkel, C. K. Chen, Finite dominating sets for network location problems, Operations Research 39 (1) (1991).

[18] E. Carrizosa, N. Mladenovic, R. Todosijevic, Variable neighborhood search for minimum sum-of-squares clustering on networks, European Journal of Operational Research 230 (2013) 356,363.

[19] C. Iyigun, A. Ben-Israel, Probabilistic distance clustering, Ph.D. thesis, Rutgers, The State University of New Jersey (2007).