JACK N. AVERITT COLLEGE of GRADUATE STUDIES GEORGIA SOUTHERN UNIVERSITY

Georgia Southern University
Digital Commons@Georgia Southern

Electronic Theses and Dissertations

Graduate Studies, Jack N. Averitt College of

Summer 2018

A Survey of Clustering Analysis and Clustering Analysis in Graphs

Raven D. Gilmore

Recommended Citation

Gilmore, Raven D., "A Survey of Clustering Analysis and Clustering Analysis in Graphs" (2018). *Electronic Theses and Dissertations*. 1781.
https://digitalcommons.georgiasouthern.edu/etd/1781

A SURVEY OF CLUSTERING ANALYSIS AND CLUSTERING ANALYSIS IN GRAPHS

by

RAVEN GILMORE

(Under the Direction of Hua Wang)

ABSTRACT

Clustering analysis is an important topic in data mining, where data points that are similar to each other are grouped together. Graph clustering deals with clustering analysis of data points that correspond to vertices on a graph. We first survey some most well known algorithms for clustering analysis. Then for graph clustering we note that one of the fundamental factors is the distance measure between vertices. We further examine various known venues for defining such measures and propose some others.

INDEX WORDS: Clustering, Graph clustering, Distance measure

2009 Mathematics Subject Classification: 90-02, 68P10

A SURVEY OF CLUSTERING ANALYSIS AND CLUSTERING ANALYSIS IN GRAPHS

by

RAVEN GILMORE

B.S., Bethune-Cookman University, 2016

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the Requirements for the Degree

.

MASTER OF SCIENCE

STATESBORO, GEORGIA

A SURVEY OF CLUSTERING ANALYSIS AND CLUSTERING ANALYSIS IN GRAPHS

by

RAVEN GILMORE

Major Professor:	Hua Wang
Committee:	Emil Iacob
	Goren Lesaja

Electronic Version Approved:
July 2018

DEDICATION

I dedicate this thesis and all of my academic achievements to my family who has been a great source of inspiration and support. This thesis is also dedicated to Dr. Hua Wang and the faculty of the department who has provided me with a constant source of knowledge and encouragement throughout my time at Georgia Southern.

ACKNOWLEDGMENTS

Thank You to:

My Family and Friends

Dr. Hua Wang

Dr. Goran Lesaja

Dr. Ionut Iacob

The Department of Mathematical Science

And to everyone who contributed to my academic success at Georgia Southern University.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION TO CLUSTERING

1.1 CLUSTERS

Clustering are the findings of a structure in a collection of unlabeled data. It is the process of organizing data into groups with members that are similar in some way, hence, forming the clusters. A cluster is the collection of data that are similar between them and are dissimilar to the data belonging to the other clusters. As shown in Figure 1.1, you can see a total of three clusters through the coloring of the dots.

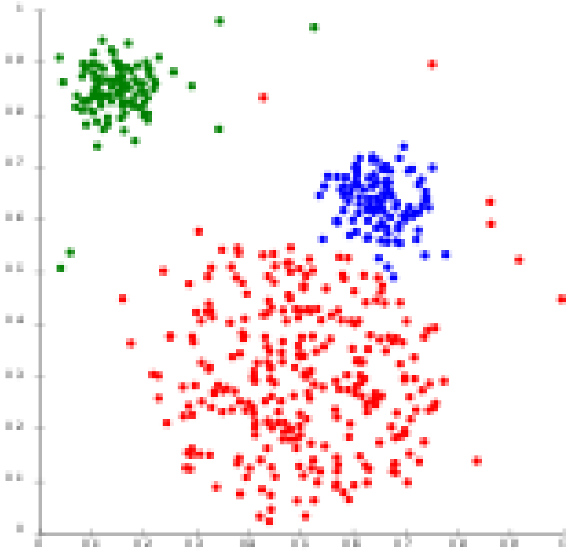

Figure 1.1: Three clusters [3].

Imagine that each dot represents a data point. Based off the euclidean distance we will measure how similar two points by how close they are to one another. As you can see in Figure 1.1 above, the given data points have been colored in three different colors. Each group of the same color generally contains points that are close to each other.

We conclude this section by presenting the formal definition of clusters.

Definition 1.1. *Clusters are groups of similar objects that are close to each other according to a certain measure of closeness.*

1.2 Why is Clustering Analysis important and what does it do?

The purpose of clustering is to make sense of the large sets of structured and unstructured data. Clustering allows you to partition that data set into logical groups before attempting to analyze it. This allows you to take a glance at all of the data, and then form a logical structure based off your findings before going deeper into more specific analysis, which is the main purpose of a clustering analysis.

Definition 1.2. *Clustering is the task of grouping objects within the same group that are more similar to one another than to those in the other groups.*

In a clustering analysis it is necessary to analyze a significant amount of information at once in regards to multiple documents. This is done by sorting, identifying, and finding both the similarities and dissimilarities between each of them. These findings are computed through several different methods, which are known as the clustering algorithms. We will briefly survey these algorithms in the next chapter.

1.3 Survey of our work

In this thesis we first present some well known clustering algorithms, together with examples through phylogenetic tree reconstruction.

We then discuss clustering analysis in graphs. More specifically, we point out that in order to conduct clustering analysis the fundamental concepts is the “distance” between data points. In graphs these data points are vetices and we are interested in vertex similarities in general. For this purpose we also present some representative distance measures defined on vertices of graphs. Such distance measures include the type based on euclidean measures and the ones induced from set differences.

Last but not least, we introduce some novel distance measures of vertex similarity, including ones that accommodate both Euclidean measures and set differences.

CHAPTER 2

CLUSTERING ALGORITHMS

Clustering algorithms consist of the different methods used to group the given data into different clusters based on their similarities. These approaches are known as the *hierarchical clustering*, *k-means clustering*, and the *two-step cluster analysis*. Each method has its own significance when it comes to performing a cluster analysis. Depending on the size of the data file it will be determined which method to use. When dealing with a large data file, it is suggested to use the two-step cluster analysis. Just as if the data file was small, then the hierarchical clustering would have been recommended to be used. Also, if you know the number of clusters you would like to have then the k-means clustering is possibly the best option.

The phylogenetic tree that is displayed Figure 2.1 shows an example of what clustering is. The tree displays a various amount of biological species and other entities bases off their similarities and differences in their physical or genetic characteristics. Intuitively speaking, species that are more alike will be grouped into closer "branches" in the phylogenetic tree.

In the rest of this chapter we will use phylogenetic tree reconstruction as an example to illustrate the three aforementioned clustering methods. A good reference that we use frequently is [1].

2.1 Hierarchical Clustering

The *hierarchical clustering*, can be used in two different ways; either the agglomerative or divisive way. Most of the time the more useful choice is the divisive way. This starts with one large cluster that contains data with a tremendous amount of files. Now based on the files and their differences, the cluster will be broken into multiple clusters. Separating those that are different and grouping those that are similar. Resulting in multiple clusters. Another way to perform the hierarchical clustering, is the agglomerative way. With this

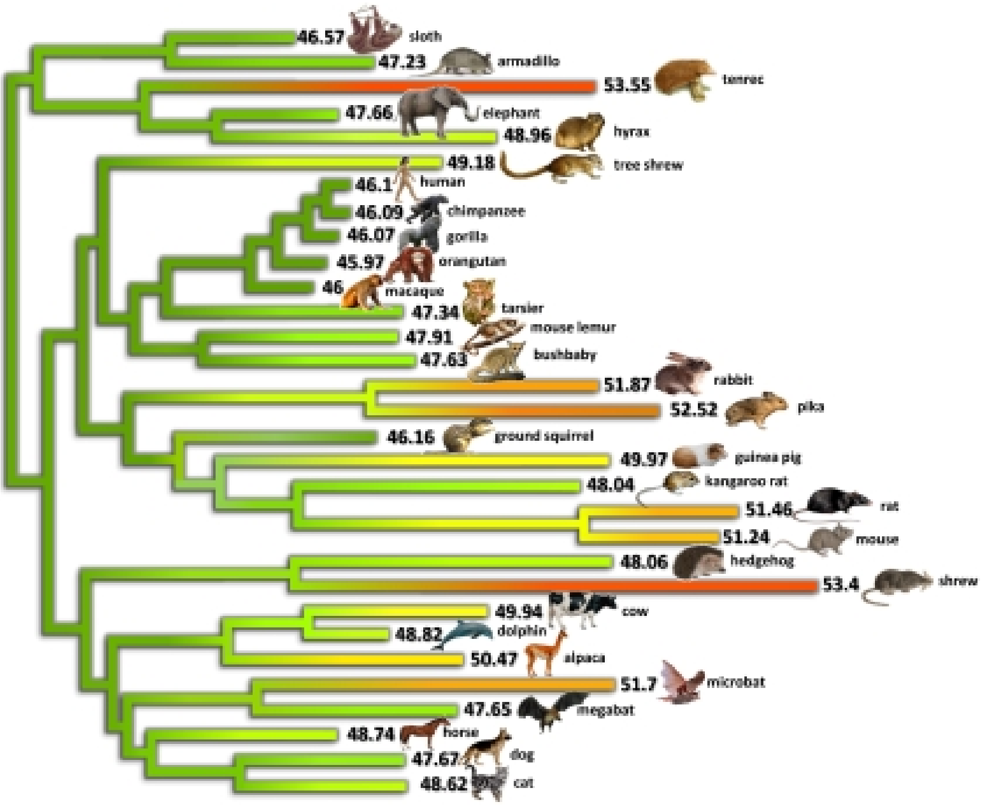

Figure 2.1: A phylogenetic tree [2].

method you are working backwards. So this means we start out with multiple clusters and end with one. This is done by finding the similarities between each group, we will continue this process until everything is grouped together, eventually giving you one big cluster [1].

So if we were to use the phylogenetic tree as an example for the *hierarchical clustering*, then the outcome will be similar to what we just explained. In Figure 2.1, we have several different types of species. Just by looking at the figure, we know if we were to use the divisive way, then, based off each species and their differences, they will be placed in their own group. The divisive method consists of 3 steps. In each step we analyze and break down everything in each group. In step one, we look at all of the species together and divide them into two groups. Group one is the Boreoeutheria species and group two are the Atlantogenata species.

Group I Boreoeutheria is a clade of placental mammals that is composed of the sister taxa Laurasiatheria and Euarchontoglires.

Group II Atlantogenata is a proposed clade of mammals containing the cohorts or superorders Afrotheria and Xenarthra.

In step two, we further analyze the differences between the species in each group. Once we have done that, we see that each group can be categorized into two sets. Group one set consist of the Euarchontoglires and Laurasiatheria species, while group two has the Afrotheria and Xenarthra species.

Group I-1 Euarchontoglires is a clade and a superorder of mammals, the living members of which belong to one of the five following groups: rodents, lagomorphs, treeshrews, colugos and primates.

Group I-2 Laurasiatheria is a superorder of placental mammals that includes shrews, pangolins, bats, whales, carnivorans, odd-toed and even-toed ungulates, among others

Group II-1 Afrotheria is a clade of mammals, the living members of which belong to groups that are either currently living in Africa or of African origin: golden moles, elephant shrews (also known as sengis), tenrecs, aardvarks, hyraxes, elephants, sea cows, and several extinct clades

Group II-2 The superorder Xenarthra is a group of placental mammals, extant today only in the Americas and represented by anteaters, tree sloths, and armadillos.

Lastly, in step three, we look at each set and separate the species based off there differences, grouping those that are similar together. Resulting in several different clusters of the species, as yo can see in the list displayed below showing the final results.

- Primates - Humans and monkeys
- Scandentia - Only tree shrews
- Rodentia - Rats, squirrels, guinea pigs, etc.
- Lagomorpha - Rabbits, hares, pikas, etc.
- Cetartiodactyla - Dolphins, cows, pigs, and alpacas
- Carnivora - Cats, lions, dogs, bears, etc.
- Perissodactyla - Horses, tapir, and rhinoceros
- Chiroptera - Bats
- Eulipotyphla - Senrecs, moles, shrews, and hedgehops.
- Proboscidea - Elephants and mastodons
- Hyracoidea - Hyraxes
- Aforsocoricida - Golden mole and tenrecs

- Cingulata - Armadillos
- Pilosa - Anteaters and sloths

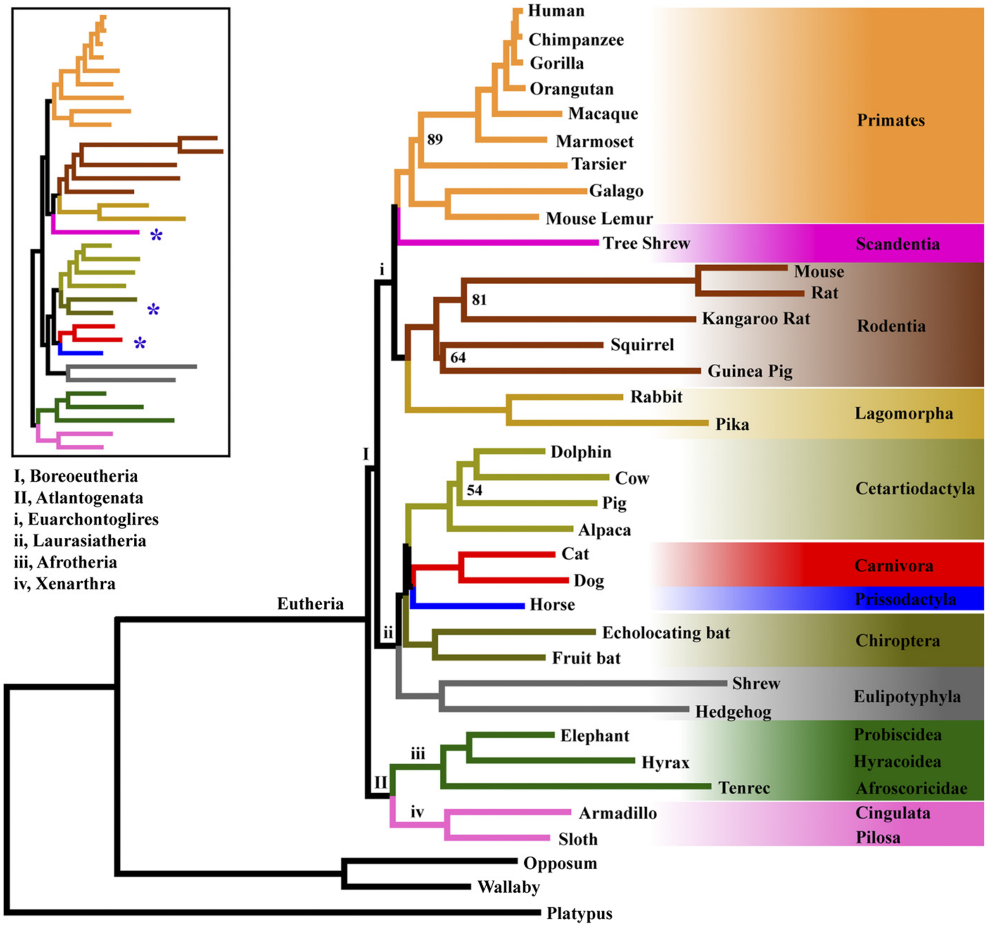

Figure 2.2: Another representation of a phylogenetic tree [10].

The second approach is the agglomerative approach. With this method we would just have to find the similarities between each species, which will eventually give us one big cluster of all the species together.

2.2 K-Means Clustering

Unlike the hierarchical clustering, the k-means clustering does not require the evaluation of all possible differences. The k-means clustering, with the k representing the number of clusters you want, starts with an initial set of means and classify cases based off their distances to the centers. Once that is done, based off the cases that were assigned to the cluster, the means must be computed again, followed by reclassifying each case based off the new set of means. This step must be repeated continuously until the cluster means does not change much between the successive steps. Lastly, the calculations of the cluster means must be done one last time, and this will result in each case being assigned to its own permanent cluster [1].

Now if we were to apply the k-means clustering method to the phylogenetic tree, then we must first decide how many clusters we want. This will be represented by k, which are the number of groups the data (species) will be placed in. Based off Figure 2.1, we known that we want a total of 14 groups (clusters). So now that we have chosen the number of cluster, we can now begin the first step of this process; the cluster assignment step. This is when the algorithm goes through each of the data points (species), and depending on which cluster is closer to that species, then it will be assigned to them. Grouping several of the species together. The next step to this method would be the "move centroid" step. This is when the algorithm calculates the average of all the points in the cluster and moves the centroid to that average location. In theory we will be repeating this process until a stopping condition is met, such as there being no change in the clusters.

2.3 Two-Step Cluster Analysis

Just as the name says, the two-step cluster analysis consists of two steps to analyze the data that is given.

The first step would be to have a formation of preclusters. The preclustering is done to reduce the size of the matrix that contains the distance between all the possible pairs of cases. The preclusters were clusters of the original cases used in place of the raw data in the hierarchical clustering. Once a case is read, the algorithm would decide, based off the distance measure, whether the current case should be merged with a previously formed precluster or start a new precluster. Once that is computed, all cases in the same precluster is treated as a single entity. This completes step one.

Now in step two, we use the standard hierarchical clustering algorithm on the preclusters. This will allow a range of solutions with different numbers of clusters, resulting in the completion of the two-step method [1].

Now, applying these steps to the phylogentic tree we have a similar outcome to what was just explained. We would conduct our first step which is to reduce the size of the matrix containing the distance between the possible pairs of species. This will allow a reading of each case, where based off the distance between each species it is possible that they will be merged together or possibly start a new precluster. Completing the first step.

Now with step two, we will apply the hierarchical clustering algorithm, where we will have a range of species with a different amount of clusters. This completes the two-step cluster on the phylogentic tree.

Sometimes this method does not seem as helpful as the other two. How well it performs largely depends on step 1 and its findings. This is because of the larger the distance measure, the fewer the preclusters you will have. But two few preclusters may result in species being grouped with other species that share nothing in common. And the smaller the distance measure, the more preclusters. However, having too many preclusters defeats the purpose of this first step.

CHAPTER 3

GRAPH CLUSTERING AND VERTEX SIMILARITIES

When thinking about clustering analysis in graphs, you must first think about a graph and what it may consist of. A graph is the structure formed by a set of vertices and a set of edges that are connections between the pairs of vertices. The graph clustering are the groupings of those vertices into clusters while taking the edges into consideration. In graph clustering, its main goal is to divide the given vertices into clusters so that each element that is assigned to a particular cluster is similar or connected in someway to other vertices.

3.1 Some graph theoretical concepts related to clustering analysis

Graph Theory is known to be the study of graphs, that are mathematical structures used to model pairwise relations between objects. In this context, a graph is made up of vertices (nodes or points) which are connected through edges (arcs or lines) (Figure 3.1). Most of the time a graph is to be considered undirected, where there is no distinction between two vertices associated with each edge. Sometimes its edges may also be directed from one vertex to another.

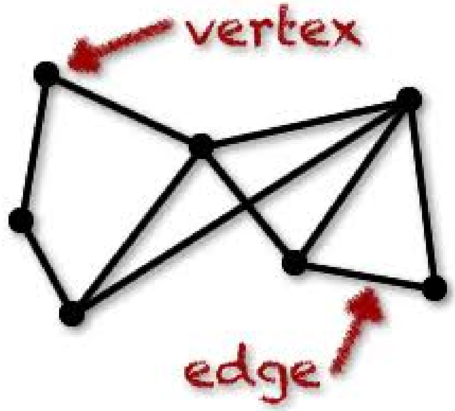

Figure 3.1: A Graph [13].

In the theory of graphs there are many examples of representations of clusters, in terms of different types of graphs, such as a subgraph or a multi-partite graph. A subgraph, is a graph G' which graph edges and vertices form a subset of the graph vertices and edges of a given graph G while the multi-partite graph is the complete k-partite graph for some k. The k-partite graph is a graph whose vertices are or can be partitioned into k different independent sets.

Similarly, there are also representations through matrices. The most well known is probably the Adjacency matrix, which displays a square matrix used to represent a finite graph. In such a matrix every entry is 0 or 1 depending on whether there exists an edge between the corresponding vertices.

Definition 3.1 (Finite Graph)**.** *A graph with a finite number of nodes and edges. If it has nodes and no multiple edges of graph loops, it is a subgraph of the complete graph.*

Definition 3.2 (Adjacency Matrix)**.** *Given a graph on n vertices, the adjacency matrix of this graph is an $n \times n$ 0-1 matrix such that the entry in the ith row and jth column is 1 if and only if there is an edge from the ith vertex to the jth vertex.*

To say a little about the history of graph clustering, we first note the introduction of random graphs.

In 1959 E.N. Gilbert [9] presented a process to generate uniform random graphs with n vertices. This allowed each of the $\binom{n}{2}$ possible edges to be included in the graph with a probability p, while considering each pair of vertices independently. In such uniform random graphs, the vertex degrees follow Poisson distribution. In addition, the presence of dense clusters are unlikely as the edges are distributed (by construction) uniformly, therefore no dense clusters can be expected.

An abstraction of Gilbert's model used to produce clusters, is the planted l-partition model [4]. The model is a graph that is generated with $n = l \cdot k$ vertices which are partitioned into groups of l with k vertices. To construct the edge set, two probability parameters p and $q < p$ are used. Each pair of vertices included in the same group share an edge with a higher probability p, whereas each pair of vertices in different groups share an edge with lower probability r. This is more or less the opposite of a multi-partite graph. The main goal of the planted partition model is to find a planted partition into l clusters of k vertices each, instead of optimizing some measure on the partition [11].

3.2 Graph Structures

Giving more of a visual, in Figure 3.2, we have an example of an adjacency matrix of a graph. On the left side the matrix is displaying n vertices and m edges. The $2m$ black dots are the ones of the matrix and the white areas are the zero entries. The left side also shows that the vertices are being ordered randomly causing the adjacency matrix to have no obvious structure, making it not suitable for interpreting the presence, number or quality of clusters inherent in the graph [12]. However, the situation changes after running a graph clustering algorithm and re-ordering the vertices according to their respective clusters. In this case we will obtain a diagonalized model of the adjacency matrix, as shown on the right side in Figure 3.2. This allows the structure of the adjacency matrix to visibly display seventeen dense clusters of different orders and some scattered connections between the clusters [11].

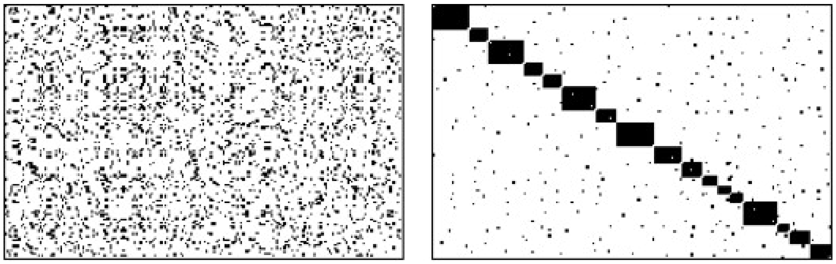

Figure 3.2: An adjacency matrix [11].

In Figure 3.3 we are shown two graphs that are of the same order and size but they are not the same graph and are identified differently. The graph on the left is a uniform random graph, where the vertices, edges and connections between them are selected in a random way, displaying a scattered graph. The graph on the right is the relaxed caveman graph. This structure is formed by linking together a ring of small complete graphs "caves" by moving one of the edges in each cave to point to another cave, which displays a clear cluster structure.

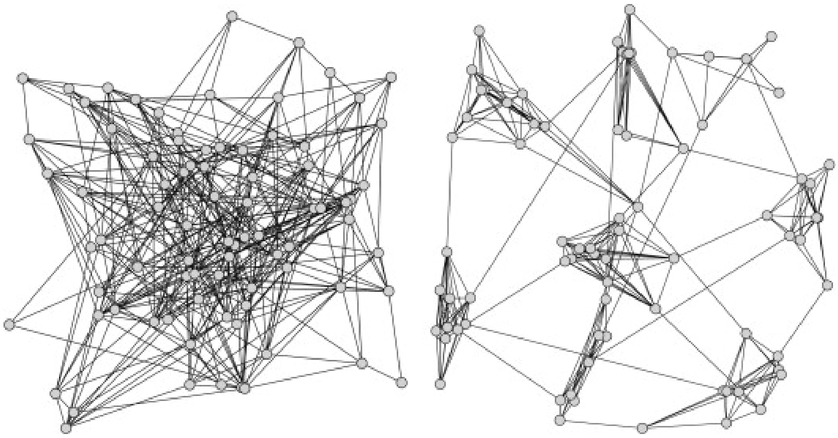

Figure 3.3: An uniform random graph and a relaxed caveman graph [6, 7, 15]

Definition 3.3 (Caveman Graph (on the right of Figure 3.3)). *A graph stemming from the social network theory that is formed by arranging a set of isolated k-cliques ("caves") by removing one edge from each clique and using it to connect to a neighboring clique along a central cycle such that all n cliques form a single unbroken loop [16].*

3.3 Vertex Similarity and Distance Measure

Performing a clustering analysis on a data file essentially depends on finding the vertex similarity. With the vertex similarities, if the vertices were being represented by documents, we would be able to compute a content-based similarity values for each of the pairs of documents, using the similarity matrix as the basis for the clustering. This is an attempt to group the vertices together that are well connected and similar to each other. Therefore, the higher the similarity, the need to cluster the vertices together becomes stronger.

If a similarity measure has been defined for vertices, then the cluster should contain vertices with close-by values and exclude those with the values differing significantly from the values of the included vertices.

Based off the task that is given, the appropriate similarity measure or distance function is chosen. From the data that is given, the distance measure should be a metric distance satisfying the following:

1. The distance from a data points to itself is zero:

$$\text{dist}(d_i, d_i) = 0$$

2. The distance are symmetrical:

$$\text{dist}(d_i, d_j) = \text{dist}(d_j, d_i)$$

3. The triangular inequality holds:

$$\text{dist}(d_i, d_j) \leq \text{dist}(d_i, d_k) + \text{dist}(d_k, d_j)$$

3.3.1 Measures based on the Euclidean distances

For points within an n-dimensional Euclidean space, common distance measure for two data points

$$d_i = (d_{i,1}, d_{i,2}, ..., d_{i,n})$$

and

$$d_j = (d_{j,1}, d_{j,2}, ..., d_{j,n})$$

include:

- the Euclidean distance:

$$\text{dist}(d_i, d_j) = \sqrt{\sum_{k=1}^{n} |d_{i,k} - d_{j,k}|^2}$$

which is the L_2 norm, also denoted as $||d_{i,k} - d_{j,k}||_2$.

Example 3.4. *Let* $d_i = (2, 4, 7, 3)$ *and* $d_j = (5, 1, 6, 8)$.

Calculate the L_2 *norm based off the given data points.*

$$|2 - 5|^2 = |-3|^2 = 9$$

$$|4 - 1|^2 = |3|^2 = 9$$

$$|7 - 6|^2 = |1|^2 = 1$$

$$|3 - 8|^2 = |-5|^2 = 25$$

$$\sqrt{9 + 9 + 1 + 25} = \sqrt{44} = 2\sqrt{11}$$

Based off the calculations, the L_2 *norm (Euclidean Distance) is* $2\sqrt{11}$.

- the Manhattan distance:

$$\text{dist}(d_i, d_j) = \sum_{k=1}^{n} |d_{i,k} - d_{j,k}|$$

which is the L_1 norm, also denoted as

$||d_{i,k} - d_{j,k}||_1$.

Example 3.5. *Let* $d_i = (2, 4, 7, 3)$ *and* $d_j = (5, 1, 6, 8)$.

Calculate the L_1 *norm based off the given data points.*

$$|2 - 5| = |-3| = 3$$

$$|4-1| = |3| = 3$$

$$|7-6| = |1| = 1$$

$$|3-8| = |-5| = 5$$

$$3+3+1+5 = 12$$

Based off the calculations, the L_1 norm (Manhattan Distance) is 12.

- and the L_∞ norm

$$\text{dist}(d_i, d_j) = \max_{k \in [1,n]} |d_{i,k} - d_{j,k}|,$$

also denoted as $||d_{i,k} - d_{j,k}||_\infty$.

Example 3.6. *Let $d_i = (2, 4, 7, 3)$ and $d_j = (5, 1, 6, 8)$.*

Calculate the L_∞ norm based off the given data points.

$$|2-5| = |-3| = 3$$

$$|4-1| = |3| = 3$$

$$|7-6| = |1| = 1$$

$$|3-8| = |-5| = 5$$

$$\max\{3, 3, 1, 5\} = 5$$

Based off the calculations, the L_∞ norm is 5.

3.3.2 Measures based on Set Distance

Very often the data points are not necessarily represented by single values or even vectors, but rather a collection of objects or values. One way to define the distance between two data points (two sets) A and B is:

$$d(A, B) = \frac{|A \triangle B|}{|A \cup B|}$$

The purpose of this formula is to compare members from the two sets to see which members are shared and which are distinct. The formula is broken down into two parts, the numerator displaying the cardinality of $|A \triangle B|$ and the denominator displaying the cardinality of $|A \cup B|$.

Definition 3.7 (Cardinality)**.** *The number of elements in a set or other grouping, as a property of that grouping.*

$A \triangle B$ or is defined as $(A - B) \cup (B - A)$ and it is the symmetric difference between sets A and B. This is the set of elements which are in either of the sets and not their intersection. As shown in (Figure 3.4), the shaded parts in the Venn diagram represents $A \triangle B$. Within those shaded regions, the number of elements will represent the cardinality of $|A \triangle B|$.

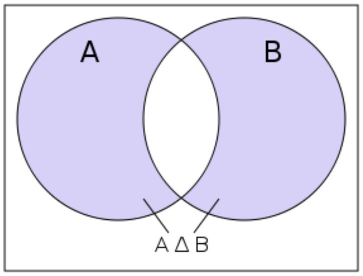

Figure 3.4: Symmetric Difference of two sets [14].

Example 3.8. *Let* $A = \{a, b, f, g, t, e, w\}$ *and* $B = \{a, c, e, y, u\}$

Determine the Symmetric Difference and its cardinality based off sets A *and* B*.*

$$A \triangle B = \{b, f, g, t, w, e, y\}$$

$$|A \triangle B| = 7$$

$A \cup B$ is the union of the two sets A and B. This is the set of elements which are in A, in B, or in both A and B. As shown in (Figure 3.5), the shaded parts in the Venn diagram represents $A \cup B$. Based off the shaded regions, the number of elements will represent the cardinality of $|A \cup B|$.

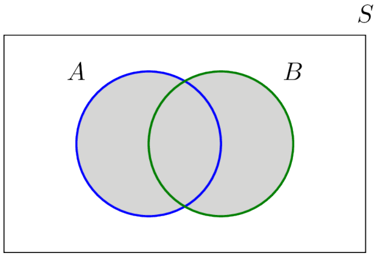

Figure 3.5: Union of two sets [9].

Example 3.9. *Let* $A = \{a, b, f, g, t, e, w\}$ *and* $B = \{a, c, e, y, u\}$

Determine the Union and its cardinality based off sets A *and* B*.*

$$A \cup B = \{a, b, c, e, f, g, t, y, w, u\}$$

$$|A \cup B| = 10$$

Hence with $A = \{a, b, f, g, t, e, w\}$ and $B = \{a, c, e, y, u\}$ we have

$$d(A, B) = \frac{|A \triangle B|}{|A \cup B|} = \frac{7}{10}.$$

Also note that this distance is never going to exceed 1 as it measures the percentage of elements in $A \cup B$ that are only in one of the two sets. Letting $C = \{a, f, g, t, e, u, w\}$, we then have

$$d(A, C) = \frac{|A \triangle C|}{|A \cup C|} = \frac{2}{9}.$$

From these simple computations we see that C is much "closer" to A than B is. This is indeed consistent with what one would have observed from their definitions.

CHAPTER 4

VERTEX SIMILARITY BASED ON THE DISTANCE METRICS AND SET DISTANCE

From what we have discussed so far, before using any established clustering algorithms it is necessary to define a distance measure between data points. In a graph, each data point is a vertex and the information for each data point comes from the structure of the graph.

In this chapter we present several direct applications of distance metrics and set distance to measure vertex similarities.

4.1 Using Distance Measures

In order to use aforementioned distance measures on vertices of a graph it is necessary to represent each vertex with a vector. The simplest way to achieve this goal is to represent the neighborhood $N(v)$ of a vertex v in terms of a 0-1 vector as follows.

For each $v \in V(G)$, let $\vec{S_v} = (\delta_1, \delta_2, ..., \delta_n)$ where $n = |V(G)|$ and

$$\delta_i = \begin{cases} 1 & \text{if } v_i \in N(v) \\ 0 & \text{if } v_i \notin N(v) \end{cases}$$

is the characteristic function.

Example 4.1. *Consider the graph G in Figure 4.1.*

$$|V(G)| = n = 4$$

There are a total of 4 vertices and 5 edges displayed on the graph above. Now based on the graph, the neighborhood of each vertex is listed as

$$N(v_1) = \{v_2, v_4\}$$

as vertex v_1 has an edge that connects with v_2 and v_4.

$$N(v_2) = \{v_1, v_3, v_4\}$$

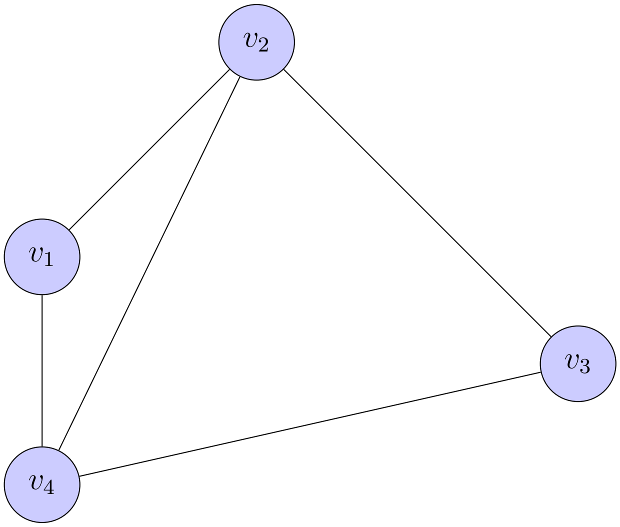

Figure 4.1: An example.

as vertex v_2 *has an edge that connects with* v_1*,* v_3*, and* v_4*.*

$$N(v_3) = \{v_2, v_4\}$$

as vertex v_3 *has an edge that connects with* v_2 *and* v_4*.*

$$N(v_4) = \{v_1, v_2, v_3\}$$

as vertex v_4 *has an edge that connects with* v_1*,* v_2*, and* v_3*.*

We now have

$$\vec{S}_{v_1} = (0, 1, 0, 1)$$

$$\vec{S}_{v_2} = (1, 0, 1, 1)$$

$$\vec{S}_{v_3} = (0, 1, 0, 1)$$

$$\vec{S}_{v_4} = (1, 1, 1, 0)$$

Note that these are exactly the rows of the adjacency matrix of G:

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

We now apply each of the metric distance based norms from the previous chapter to S_{v_i}'s.

- L_2 Norm (Euclidean Distance)

  Recall that

  $$\text{dist}(d_i, d_j) = \sqrt{\sum_{k=1}^{n} |d_{i,k} - d_{j,k}|^2},$$

  then

  $$\begin{aligned} \text{dist}(S_{v_1}, S_{v_2}) &= \sqrt{|0-1|^2 + |1-0|^2 + |0-1|^2 + |1-1^2|} \\ &= \sqrt{|-1|^2 + |1|^2 + |-1|^2 + |0|^2} \\ &= \sqrt{1+1+1+0} \\ &= \sqrt{3} \end{aligned}$$

  $$\begin{aligned} \text{dist}(S_{v_1}, S_{v_3}) &= \sqrt{|0-0|^2 + |1-1|^2 + |0-0|^2 + |1-1^2|} \\ &= \sqrt{|0|^2 + |0|^2 + |0|^2 + |0|^2} \\ &= \sqrt{0+0+0+0} \\ &= \sqrt{0} \\ &= 0 \end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_1}, S_{v_4}) =& \sqrt{|0-1|^2 + |1-1|^2 + |0-1|^2 + |1-0^2|}\\ =& \sqrt{|-1|^2 + |0|^2 + |-1|^2 + |1|^2}\\ =& \sqrt{1+0+1+1}\\ =& \sqrt{3}\end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_2}, S_{v_3}) =& \sqrt{|1-0|^2 + |0-1|^2 + |1-0|^2 + |1-1^2|}\\ =& \sqrt{|1|^2 + |-1|^2 + |1|^2 + |0|^2}\\ =& \sqrt{1+1+1+0}\\ =& \sqrt{3}\end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_2}, S_{v_4}) =& \sqrt{|1-1|^2 + |0-1|^2 + |1-1|^2 + |1-0^2|}\\ =& \sqrt{|0|^2 + |-1|^2 + |0|^2 + |1|^2}\\ =& \sqrt{0+1+0+1}\\ =& \sqrt{2}\end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_3}, S_{v_4}) =& \sqrt{|0-1|^2 + |1-1|^2 + |0-1|^2 + |1-0^2|}\\ =& \sqrt{|-1|^2 + |0|^2 + |-1|^2 + |1|^2}\\ =& \sqrt{1+0+1+1}\\ =& \sqrt{3}\end{aligned}$$

Notice that S_{v_1} and S_{v_3} are exactly the same because of the identical neighborhood of the corresponding vertices. Consequently the distance between these two will be zero.

- L_1 Norm (Manhattan Distance)

  Recall that

  $$\text{dist}(d_i, d_j) = \sum_{k=1}^{n} |d_{i,k} - d_{j,k}|,$$

  then

$$\begin{aligned}\text{dist}(S_{v_1}, S_{v_2}) &= |0-1|^2 + |1-0|^2 + |0-1|^2 + |1-1^2| \\ &= |-1|^2 + |1|^2 + |-1|^2 + |0|^2 \\ &= 1+1+1+0 \\ &= 3\end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_1}, S_{v_3}) &= |0-0|^2 + |1-1|^2 + |0-0|^2 + |1-1^2| \\ &= |0|^2 + |0|^2 + |0|^2 + |0|^2 \\ &= 0+0+0+0 \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_1}, S_{v_4}) &= |0-1|^2 + |1-1|^2 + |0-1|^2 + |1-0^2| \\ &= |-1|^2 + |0|^2 + |-1|^2 + |1|^2 \\ &= 1+0+1+1 \\ &= 3\end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_2}, S_{v_3}) &= |1-0|^2 + |0-1|^2 + |1-0|^2 + |1-1^2| \\ &= |1|^2 + |-1|^2 + |1|^2 + |0|^2 \\ &= 1+1+1+0 \\ &= 3\end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_2}, S_{v_4}) &= |1-1|^2 + |0-1|^2 + |1-1|^2 + |1-0^2| \\ &= |0|^2 + |-1|^2 + |0|^2 + |1|^2 \\ &= 0+1+0+1 \\ &= 2\end{aligned}$$

$$\begin{aligned}\text{dist}(S_{v_3}, S_{v_4}) &= |0-1|^2 + |1-1|^2 + |0-1|^2 + |1-0^2| \\ &= |-1|^2 + |0|^2 + |-1|^2 + |1|^2 \\ &= 1+0+1+1 \\ &= 3\end{aligned}$$

- L_∞ Norm (Chebyshev Distance)

  Recall that

  $$\text{dist}(d_i, d_j) = \max_{k\in[1,n]} |d_{i,k} - d_{j,k}|,$$

  then

$\text{max}_{k\in[4]} \lvert S_{v_1} - S_{v_2} \rvert$	$\text{max}_{k\in[4]} \lvert S_{v_1} - S_{v_3} \rvert$
$\lvert 0-1 \rvert = 1$	$\lvert 0-0 \rvert = 0$
$\lvert 1-0 \rvert = 1$	$\lvert 1-1 \rvert = 0$

$|0-1| = 1$

$|1-1| = 0$

$\max_{k\in[4]}\{1,1,1,0\} = 1$

$|0-0| = 0$

$|1-1| = 0$

$\max_{k\in[4]}\{0,0,0,0\} = 0$

$\max_{k\in[4]} |S_{v_1} - S_{v_4}|$

$|0-1| = 1$

$|1-1| = 0$

$|0-1| = 1$

$|1-0| = 1$

$\max_{k\in[4]}\{1,0,1,1\} = 1$

$\max_{k\in[4]} |S_{v_2} - S_{v_3}|$

$|1-0| = 1$

$|0-1| = 1$

$|1-0| = 1$

$|1-1| = 0$

$\max_{k\in[4]}\{1,1,1,0\} = 1$

$\max_{k\in[4]} |S_{v_2} - S_{v_4}|$

$|1-1| = 0$

$|0-1| = 1$

$|1-1| = 0$

$|1-0| = 1$

$\max_{k\in[4]}\{0,1,0,1\} = 1$

$\max_{k\in[4]} |S_{v_3} - S_{v_4}|$

$|0-1| = 1$

$|1-1| = 0$

$|0-1| = 1$

$|1-0| = 1$

$\max_{k\in[4]}\{1,0,1,1\} = 1$

4.2 USING SET DISTANCE

Similarly, in order to use set distance to measure vertex similarity we need to represent each vertex with a set. The natural approach is to use a vertex's neighborhood $N(v)$ for each $v \in V(G)$.

Example 4.2. *Again using the graph G from Figure 4.1, we have:*

$$N(v_1) = \{v_2, v_4\}$$

$$N(v_2) = \{v_1, v_3, v_4\}$$

$$N(v_3) = \{v_2, v_4\}$$

$$N(v_4) = \{v_1, v_2, v_3\}$$

Recall that

$$d(A, B) = \frac{|A \triangle B|}{|A \cup B|},$$

applying this formula to $N(v_i)$'s yields the following.

$$\begin{aligned} d(N(v_1), N(v_2)) &= \frac{|N(v_1) \triangle N(v_2)|}{|N(v_1) \cup N(v_2)|} \\ &= \frac{|\{v_1, v_2, v_3\}|}{|\{v_1, v_2, v_3, v_4\}|} \\ &= \frac{3}{4} \end{aligned}$$

$$\begin{aligned} d(N(v_1), N(v_3)) &= \frac{|N(v_1) \triangle N(v_3)|}{|N(v_1) \cup N(v_3)|} \\ &= \frac{|\emptyset|}{|v_2, v_4|} \\ &= \frac{0}{2} \\ &= 0 \end{aligned}$$

$$\begin{aligned} d(N(v_1), N(v_4)) &= \frac{|N(v_1) \triangle N(v_4)|}{|N(v_1) \cup N(v_4)|} \\ &= \frac{|\{v_1, v_3, v_4\}|}{|\{v_1, v_2, v_3, v_4\}|} \\ &= \frac{3}{4} \end{aligned}$$

$$\begin{aligned} d(N(v_2), N(v_3)) &= \frac{|N(v_2) \triangle N(v_3)|}{|N(v_2) \cup N(v_3)|} \\ &= \frac{|\{v_1, v_2, v_3\}|}{|\{v_1, v_2, v_3, v_4\}|} \\ &= \frac{3}{4} \end{aligned}$$

$$\begin{aligned} d(N(v_2), N(v_4)) &= \frac{|N(v_2) \triangle N(v_4)|}{|N(v_2) \cup N(v_4)|} \\ &= \frac{|\{v_2, v_4\}|}{|\{v_1, v_2, v_3, v_4\}|} \\ &= \frac{2}{4} \\ &= \frac{1}{2} \end{aligned}$$

$$\begin{aligned} d(N(v_3), N(v_4)) &= \frac{|N(v_3) \triangle N(v_4)|}{|N(v_3) \cup N(v_4)|} \\ &= \frac{|\{v_1, v_3, v_4\}|}{|\{v_1, v_2, v_3, v_4\}|} \\ &= \frac{3}{4} \end{aligned}$$

Example 4.3. *It is also interesting to apply set difference to the $S_(v_i)'s$ as sets instead of vectors.*

$$\begin{aligned} d(S_{v_1}, S_{v_2}) &= \frac{|S_{v_1} \triangle S_{v_2}|}{|S_{v_1} \cup S_{v_2}|} \\ &= \frac{3}{5} \end{aligned} \qquad \begin{aligned} d(S_{v_1}, S_{v_3}) &= \frac{|S_{v_1} \triangle S_{v_3}|}{|S_{v_1} \cup S_{v_3}|} \\ &= \frac{0}{4} \\ &= 0 \end{aligned}$$

$$\begin{aligned} d(S_{v_1}, S_{v_4}) &= \frac{|S_{v_1} \triangle S_{v_4}|}{|S_{v_1} \cup S_{v_4}|} \\ &= \frac{3}{5} \end{aligned} \qquad \begin{aligned} d(S_{v_2}, S_{v_3}) &= \frac{|S_{v_2} \triangle S_{v_3}|}{|S_{v_2} \cup S_{v_3}|} \\ &= \frac{3}{5} \end{aligned}$$

$$d(S_{v_2}, S_{v_4}) = \frac{|S_{v_2} \triangle S_{v_4}|}{|S_{v_2} \cup S_{v_4}|} = \frac{2}{6} = \frac{1}{3} \qquad d(S_{v_3}, S_{v_4}) = \frac{|S_{v3} \triangle S_{v_4}|}{|S_{v_3} \cup S_{v_4}|} = \frac{3}{5}$$

4.3 Comparison of different measures

Before ending this chapter, we compare the resulted vertex similarities obtained using different measures. Table 4.1 displays the results of each Distance Metrics and Set Distance.

Table 4.1: Comparison of vertex similarities under different measures.

	L_2	L_1	L_∞	Set Diff.
$d(v_1, v_2)$	$\sqrt{3}$	3	1	$\frac{3}{4}$
$d(v_1, v_3)$	0	0	0	0
$d(v_1, v_4)$	$\sqrt{3}$	3	1	$\frac{3}{4}$
$d(v_2, v_3)$	$\sqrt{3}$	3	1	$\frac{3}{4}$
$d(v_2, v_4)$	$\sqrt{2}$	2	1	$\frac{1}{2}$
$d(v_3, v_4)$	$\sqrt{3}$	3	1	$\frac{3}{4}$

As expected, for vertices v_1 and v_3 with identical neighborhood. The "distance" between them is always zero regardless of the measures that are employed. Between the other vertices, v_2 and v_4, although different from each other, shows more similarity (i.e. less difference) than other pairs of vertices. Regardless of the euclidean distance or set distance measures that were used our results appear to be consistent.

CHAPTER 5

NOVEL VERTEX SIMILARITY MEASURES WITH EXAMPLES

As one can see from the previous chapter, a particular distance measure or representation of vertex information is very often not sufficient to distinguish the difference between different pairs of vertices. For instance, in Table 4.1, several different pair of vertices share the same value when using the few basic distance measures.

It is then natural to include more information when representing vertices from a graph and develop slightly more sensitive distance measures. In this chapter we will introduce two representations of vertices that accommodate not only the neighborhoods, but the collections of vertices at distance i for any i.

For each of the proposed model we show, in detail, how to evaluate the differences. We also apply them to an exemplary graph (Figure 5.1) and compare the results.

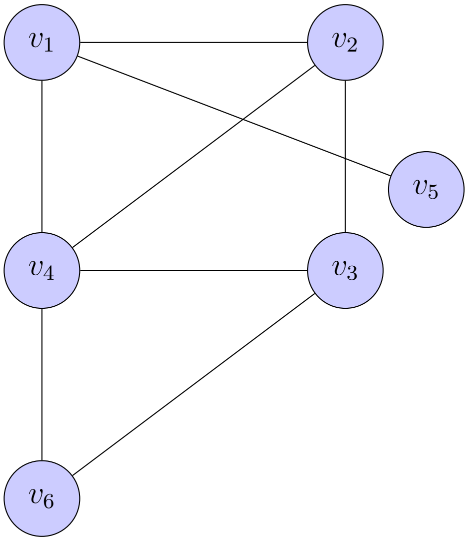

Figure 5.1: Another example.

5.1 VECTOR REPRESENTATION AND DISTANCE MEASURES

First let

$$N_i(v) = \{u \in V(G) | d_G(u, v) = i\}.$$

This is the set of vertices that are at distance i from v. That is, the i is the number of edges it takes to connected v to any vetex in that set.

Then

$$\delta_i(v) = |N_i(v)|$$

is simply the number of vertices at distance i from v, for each vertex v of the graph G.

We now define

$$\vec{R}_v = \{(\delta_1(v), \delta_2(v), \delta_3(v), ..., \delta_k(v))\} \text{ where } k = diam(G),$$

the vector representation of a vertex v based on the number of vertices at various distances from v. In other words, it is a vector keeping track of how many times a distance occur from a particular vertex. Since the maximum such distance in G is called the diameter $diam(G) = k$, we only need to consider the values of i up to k.

First we list the distances between each pair of vertices:

$$\begin{array}{ll}
d_G(v_1, v_2) = 1 & d_G(v_1, v_3) = 2 \\
d_G(v_1, v_4) = 1 & d_G(v_1, v_5) = 1 \\
d_G(v_1, v_6) = 2 & d_G(v_2, v_3) = 1 \\
d_G(v_2, v_4) = 1 & d_G(v_2, v_5) = 2 \\
d_G(v_2, v_6) = 2 & d_G(v_3, v_4) = 1 \\
d_G(v_3, v_5) = 3 & d_G(v_3, v_6) = 1 \\
d_G(v_4, v_5) = 2 & d_G(v_4, v_6) = 1 \\
d_G(v_5, v_6) = 3 &
\end{array}$$

Now we obtain $N_s(v_i)$ for any $1 \leq s \leq 3$ and $1 \leq i \leq 6$ in Figure 5.1.

Distance of 1	**Distance of 2**	**Distance of 3**
$N_1(v_1) = \{v_2, v_4, v_5\}$	$N_2(v_1) = \{v_3, v_6\}$	$N_3(v_1) = \emptyset$
$N_1(v_2) = \{v_1, v_3, v_4\}$	$N_2(v_2) = \{v_5, v_6\}$	$N_3(v_2) = \emptyset$
$N_1(v_3) = \{v_2, v_4, v_6\}$	$N_2(v_3) = \{v_1\}$	$N_3(v_3) = \{v_5\}$
$N_1(v_4) = \{v_1, v_2, v_3, v_6\}$	$N_2(v_4) = \{v_5\}$	$N_3(v_4) = \emptyset$
$N_1(v_5) = \{v_1\}$	$N_2(v_5) = \{v_2, v_4\}$	$N_3(v_5) = \{v_3, v_6\}$
$N_1(v_6) = \{v_3, v_4\}$	$N_2(v_6) = \{v_1, v_2\}$	$N_3(v_6) = \{v_5\}$

By taking the cardinalities of the above sets we now have the values of $\delta_s(v_i)$:

$\delta_1(v_1) = 3$	$\delta_2(v_1) = 2$	$\delta_3(v_1) = 0$
$\delta_1(v_2) = 3$	$\delta_2(v_2) = 2$	$\delta_3(v_2) = 0$
$\delta_1(v_3) = 3$	$\delta_2(v_3) = 1$	$\delta_3(v_3) = 1$
$\delta_1(v_4) = 4$	$\delta_2(v_4) = 1$	$\delta_3(v_4) = 0$
$\delta_1(v_5) = 1$	$\delta_2(v_5) = 2$	$\delta_3(v_5) = 2$
$\delta_1(v_6) = 2$	$\delta_2(v_6) = 2$	$\delta_3(v_6) = 1$

Now, applying our model to Figure 5.1, we have:

- $\vec{R}_{v_1}$ = (3,2,0) from

$$N_1(v_1) = \{v_2, v_4, v_5\} \rightarrow \delta_1(v_1) = 3$$

$$N_2(v_1) = \{v_3, v_6\} \rightarrow \delta_2(v_1) = 2$$

$$N_3(v_1) = \emptyset \rightarrow \delta_3(v_1) = 0$$

- $\vec{R}_{v_2}$ = (3,2,0) from

$$N_1(v_2) = \{v_1, v_3, v_4\} \rightarrow \delta_1(v_2) = 3$$
$$N_2(v_2) = \{v_5, v_6\} \rightarrow \delta_2(v_2) = 2$$
$$N_3(v_2) = \emptyset \rightarrow \delta_3(v_2) = 0$$

- $\vec{R}_{v_3}$ = (3,1,1) from

$$N_1(v_3) = \{v_2, v_4, v_6\} \rightarrow \delta_1(v_3) = 3$$
$$N_2(v_3) = \{v_1\} \rightarrow \delta_2(v_3) = 1$$
$$N_3(v_3) = \{v_5\} \rightarrow \delta_3(v_3) = 1$$

$\vec{R}_{v_4}$ = (4,1,0) from

$$N_1(v_4) = \{v_1, v_2, v_3, v_6\} \rightarrow \delta_1(v_4) = 4$$
$$N_2(v_4) = \{v_5\} \rightarrow \delta_2(v_4) = 1$$
$$N_3(v_4) = \emptyset \rightarrow \delta_3(v_4) = 0$$

- $\vec{R}_{v_5}$ = (1,2,2) from

$$N_1(v_5) = \{v_1\} \rightarrow \delta_1(v_5) = 1$$
$$N_2(v_5) = \{v_2, v_4\} \rightarrow \delta_2(v_5) = 2$$
$$N_3(v_5) = \{v_3, v_6\} \rightarrow \delta_3(v_5) = 2$$

- $\vec{R}_{v_6}$ = (2,2,1) from

$$N_1(v_6) = \{v_3, v_4\} \rightarrow \delta_1(v_6) = 2$$
$$N_2(v_6) = \{v_1, v_2\} \rightarrow \delta_2(v_6) = 2$$
$$N_3(v_6) = \{v_5\} \rightarrow \delta_3(v_6) = 1$$

It may be interesting to note that, the sum of the values in each vector is exactly the total number of other vertices in the graph (i.e. $|V(G)| - 1$). In the figure above there are a total of six vertices, and

$$\vec{R}_{v_1} = (3,2,0) \rightarrow 3 + 2 + 0 = 5$$
$$\vec{R}_{v_1} = (3,2,0) \rightarrow 3 + 2 + 0 = 5$$
$$\vec{R}_{v_3} = (3,1,1) \rightarrow 3 + 1 + 1 = 5$$
$$\vec{R}_{v_4} = (4,1,0) \rightarrow 4 + 1 + 0 = 5$$
$$\vec{R}_{v_5} = (1,2,2) \rightarrow 1 + 2 + 2 = 5$$
$$\vec{R}_{v_6} = (2,2,1) \rightarrow 2 + 2 + 1 = 5$$

Now we have:

$$\vec{R}_{v_1} = (3, 2, 0)$$
$$\vec{R}_{v_1} = (3, 2, 0)$$
$$\vec{R}_{v_3} = (3, 1, 1)$$
$$\vec{R}_{v_4} = (4, 1, 0)$$
$$\vec{R}_{v_5} = (1, 2, 2)$$
$$\vec{R}_{v_6} = (2, 2, 1)$$

Applying L_2 norm to each pair of vectors, we have

$$d(\vec{R}_{v_i}, \vec{R}_{v_j}) = \sqrt{\sum_{s=1}^{k} (\delta_s(v_i) - \delta_s(v_j))^2}.$$

Then

$$d(\vec{R}_{v_1},\vec{R}_{v_2}) = \sqrt{(3-3)^2+(2-2)^2+(0-0)^2} = \sqrt{0} = 0$$
$$d(\vec{R}_{v_1},\vec{R}_{v_3}) = \sqrt{(3-3)^2+(2-1)^2+(0-1)^2} = \sqrt{1+1} = \sqrt{2}$$
$$d(\vec{R}_{v_1},\vec{R}_{v_4}) = \sqrt{(3-4)^2+(2-1)^2+(0-0)^2} = \sqrt{1+1} = \sqrt{2}$$
$$d(\vec{R}_{v_1},\vec{R}_{v_5}) = \sqrt{(3-1)^2+(2-2)^2+(0-2)^2} = \sqrt{4+4} = 2\sqrt{2}$$
$$d(\vec{R}_{v_1},\vec{R}_{v_6}) = \sqrt{(3-2)^2+(2-2)^2+(0-1)^2} = \sqrt{1+1} = \sqrt{2}$$

$$d(\vec{R}_{v_2},\vec{R}_{v_3}) = \sqrt{(3-3)^2+(2-1)^2+(0-1)^2} = \sqrt{1+1} = \sqrt{2}$$
$$d(\vec{R}_{v_2},\vec{R}_{v_4}) = \sqrt{(3-4)^2+(2-1)^2+(0-0)^2} = \sqrt{1+1} = \sqrt{2}$$
$$d(\vec{R}_{v_2},\vec{R}_{v_5}) = \sqrt{(3-1)^2+(2-2)^2+(0-2)^2} = \sqrt{4+4} = 2\sqrt{2}$$
$$d(\vec{R}_{v_2},\vec{R}_{v_6}) = \sqrt{(3-2)^2+(2-2)^2+(0-1)^2} = \sqrt{1+1} = \sqrt{2}$$

$$d(\vec{R}_{v_3},\vec{R}_{v_4}) = \sqrt{(3-4)^2+(1-1)^2+(1-0)^2} = \sqrt{1+1} = \sqrt{2}$$
$$d(\vec{R}_{v_3},\vec{R}_{v_5}) = \sqrt{(3-1)^2+(1-2)^2+(1-2)^2} = \sqrt{4+1+1} = \sqrt{6}$$
$$d(\vec{R}_{v_3},\vec{R}_{v_6}) = \sqrt{(3-2)^2+(1-2)^2+(1-1)^2} = \sqrt{1+1} = \sqrt{2}$$

$$d(\vec{R}_{v_4},\vec{R}_{v_5}) = \sqrt{(4-1)^2+(1-2)^2+(0-2)^2} = \sqrt{9+1+4} = \sqrt{14}$$
$$d(\vec{R}_{v_4},\vec{R}_{v_6}) = \sqrt{(4-2)^2+(1-2)^2+(0-1)^2} = \sqrt{4+1+1} = \sqrt{6}$$

$$d(\vec{R}_{v_5},\vec{R}_{v_6}) = \sqrt{(1-2)^2+(2-2)^2+(2-1)^2} = \sqrt{1+1} = \sqrt{2}$$

Similarly, applying L_1 Norm and L_∞ Norm yield

$$d(\vec{R}_{v_i},\vec{R}_{v_j}) = \sum_{s=1}^{k} |\delta_s(v_i) - \delta_s(v_j)|$$

and

$$d(\vec{R}_{v_i}, \vec{R}_{v_j}) = \max_{s \in [k]} |\delta_s(v_i) - \delta_s(v_j)|,$$

respectively. We skip the details but summarize our findings in Table 5.1. One can easily see a larger variety of different distances between different pairs of vertices than in the previous chapter.

Table 5.1: Distance Metrics between each $\vec{R}_{v_i}$'s.

	$L_2\ Norm$	$L_1\ Norm$	$L_\infty\ Norm$
$d(\vec{R}_{v_1}, \vec{R}_{v_2})$	0	0	0
$d(\vec{R}_{v_1}, \vec{R}_{v_3})$	$\sqrt{2}$	2	1
$d(\vec{R}_{v_1}, \vec{R}_{v_4})$	$\sqrt{2}$	2	1
$d(\vec{R}_{v_1}, \vec{R}_{v_5})$	$2\sqrt{2}$	4	2
$d(\vec{R}_{v_1}, \vec{R}_{v_6})$	$\sqrt{2}$	2	1
$d(\vec{R}_{v_2}, \vec{R}_{v_3})$	$\sqrt{2}$	2	1
$d(\vec{R}_{v_2}, \vec{R}_{v_4})$	$\sqrt{2}$	2	1
$d(\vec{R}_{v_2}, \vec{R}_{v_5})$	$2\sqrt{2}$	4	2
$d(\vec{R}_{v_2}, \vec{R}_{v_6})$	$\sqrt{2}$	2	1
$d(\vec{R}_{v_3}, \vec{R}_{v_4})$	$\sqrt{2}$	2	1
$d(\vec{R}_{v_3}, \vec{R}_{v_5})$	$\sqrt{6}$	4	2
$d(\vec{R}_{v_3}, \vec{R}_{v_6})$	$\sqrt{2}$	2	1
$d(\vec{R}_{v_4}, \vec{R}_{v_5})$	$\sqrt{14}$	6	3
$d(\vec{R}_{v_4}, \vec{R}_{v_6})$	$\sqrt{6}$	4	2
$d(\vec{R}_{v_5}, \vec{R}_{v_6})$	$\sqrt{2}$	2	1

5.2 Vector representation and set distance

This time for each vertex v we define

$$\vec{W}_v = (N_1(v), N_2(v), ..., N_k(v)).$$

That is, instead of just tracking the number of vertices at various distances from v we record the set of these vertices. Again from Figure 5.1 we have

$$\begin{array}{lll}
N_1(v_1) = \{v_2, v_4, v_5\} & N_2(v_1) = \{v_3, v_6\} & N_3(v_1) = \emptyset \\
N_1(v_2) = \{v_1, v_3, v_4\} & N_2(v_2) = \{v_5, v_6\} & N_3(v_2) = \emptyset \\
N_1(v_3) = \{v_2, v_4, v_6\} & N_2(v_3) = \{v_1\} & N_3(v_3) = \{v_5\} \\
N_1(v_4) = \{v_1, v_2, v_3, v_6\} & N_2(v_4) = \{v_5\} & N_3(v_4) = \emptyset \\
N_1(v_5) = \{v_1\} & N_2(v_5) = \{v_2, v_4\} & N_3(v_5) = \{v_3, v_6\} \\
N_1(v_6) = \{v_3, v_4\} & N_2(v_6) = \{v_1, v_2\} & N_3(v_6) = \{v_5\}
\end{array}$$

Hence

$$\vec{W}_{v_1} = (\{v_2, v_4, v_5\}, \{v_3, v_6\}, \emptyset)$$

$$\vec{W}_{v_2} = (\{v_1, v_3, v_4\}, \{v_5, v_6\}, \emptyset)$$

$$\vec{W}_{v_3} = (\{v_2, v_4, v_6\}, \{v_1\}, \{v_5\})$$

$$\vec{W}_{v_4} = (\{v_1, v_2, v_3., v_6\}, \{v_5\}, \emptyset)$$

$$\vec{W}_{v_5} = (\{v_1\}, \{v_2, v_4\}, \{v_3, v_6\})$$

$$\vec{W}_{v_6} = (\{v_3, v_4\}, \{v_1, v_2\}, \{v_5\})$$

Recall now the difference formula for a pair of sets A and B:

$$d(A, B) = \frac{|A \triangle B|}{|A \cup B|},$$

we have

Table 5.2: Set Distance between each W_{v_i}'s.

	$i=1$	$i=2$	$i=3$
$d(\vec{W}_{v_1}, \vec{W}_{v_2})$	$\frac{4}{5}$	$\frac{2}{3}$	0
$d(\vec{W}_{v_1}, \vec{W}_{v_3})$	$\frac{1}{2}$	1	1
$d(\vec{W}_{v_1}, \vec{W}_{v_4})$	$\frac{5}{6}$	1	0
$d(\vec{W}_{v_1}, \vec{W}_{v_5})$	1	1	1
$d(\vec{W}_{v_1}, \vec{W}_{v_6})$	$\frac{3}{4}$	1	1
$d(\vec{W}_{v_2}, \vec{W}_{v_3})$	$\frac{4}{5}$	1	1
$d(\vec{W}_{v_2}, \vec{W}_{v_4})$	$\frac{3}{5}$	$\frac{1}{2}$	0
$d(\vec{W}_{v_2}, \vec{W}_{v_5})$	$\frac{2}{3}$	1	1
$d(\vec{W}_{v_2}, \vec{W}_{v_6})$	$\frac{1}{3}$	1	1
$d(\vec{W}_{v_3}, \vec{W}_{v_4})$	$\frac{3}{5}$	1	1
$d(\vec{W}_{v_3}, \vec{W}_{v_5})$	1	1	1
$d(\vec{W}_{v_3}, \vec{W}_{v_6})$	$\frac{3}{4}$	$\frac{1}{2}$	0
$d(\vec{W}_{v_4}, \vec{W}_{v_5})$	$\frac{3}{4}$	1	1
$d(\vec{W}_{v_4}, \vec{W}_{v_6})$	$\frac{4}{5}$	1	1
$d(\vec{W}_{v_5}, \vec{W}_{v_6})$	1	$\frac{2}{3}$	1

By applying L_2 Norm to the numerical valued vectors as a result of taking corresponding set distance, we have

$$d(\vec{W}_{v_i}, \vec{W}_{v_j}) = \sqrt{\sum_{s=1}^{k} [d(N_s(v_i), N_s(v_j))]^2}.$$

Conducting this computation for all pairs of vertices, we have the following.

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_2}) &= \sqrt{\left(\frac{4}{5}\right)^2 + \left(\frac{2}{3}\right)^2 + (0)^2} \\ &= \sqrt{\frac{244}{225}} \\ &= \frac{2\sqrt{61}}{15} \\ &\approx 1.04 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_3}) &= \sqrt{\left(\frac{1}{2}\right)^2 + (1)^2 + (1)^2} \\ &= \sqrt{\frac{9}{4}} \\ &= \frac{3}{2} \\ &\approx 1.5 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_4}) &= \sqrt{\left(\frac{5}{6}\right)^2 + (1)^2 + (0)^2} \\ &= \sqrt{\frac{61}{36}} \\ &= \frac{\sqrt{61}}{6} \\ &\approx 1.30 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_5}) &= \sqrt{(1)^2 + (1)^2 + (1)^2} \\ &= \sqrt{3} \\ &\approx 1.73 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_6}) &= \sqrt{\left(\frac{3}{4}\right)^2 + (1)^2 + (1)^2} \\ &= \sqrt{\frac{41}{16}} \\ &= \frac{\sqrt{41}}{4} \\ &\approx 1.60 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_2}, \vec{W}_{v_3}) &= \sqrt{\left(\frac{4}{5}\right)^2 + (1)^{+}(1)^2} \\ &= \sqrt{\frac{66}{25}} \\ &= \frac{\sqrt{66}}{5} \\ &\approx 1.62 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_2}, \vec{W}_{v_4}) &= \sqrt{\left(\frac{3}{5}\right)^2 + \left(\frac{1}{2}\right)^2 + (0)^2} \\ &= \sqrt{\frac{61}{100}} \\ &= \frac{\sqrt{61}}{10} \\ &\approx 0.78 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_2}, \vec{W}_{v_5}) &= \sqrt{\left(\frac{2}{3}\right)^2 + (1)^2 + (1)^2} \\ &= \sqrt{\frac{22}{9}} \\ &= \frac{\sqrt{22}}{3} \\ &\approx 1.56 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_2}, \vec{W}_{v_6}) &= \sqrt{\left(\frac{1}{3}\right)^2 + (1)^2 + (1)^2} \\ &= \sqrt{\frac{19}{9}} \\ &= \frac{\sqrt{19}}{3} \\ &\approx 1.45 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_3}, \vec{W}_{v_4}) &= \sqrt{\left(\frac{3}{5}\right)^2 + (1)^2 + (1)^2} \\ &= \sqrt{\frac{59}{25}} \\ &= \frac{\sqrt{59}}{5} \\ &\approx 1.54 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_3}, \vec{W}_{v_5}) &= \sqrt{(1)^2 + (1)^2 + (1)^2} \\ &= \sqrt{3} \\ &\approx 1.73 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_3}, \vec{W}_{v_6}) &= \sqrt{\left(\frac{3}{4}\right)^2 + \left(\frac{1}{2}\right)^2 + (0)^2} \\ &= \sqrt{\frac{13}{16}} \\ &= \frac{\sqrt{13}}{4} \\ &\approx 0.9 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_4}, \vec{W}_{v_5}) &= \sqrt{\left(\frac{3}{4}\right)^2 + (1)^2 + (1)^2} \\ &= \sqrt{\frac{41}{16}} \\ &= \frac{\sqrt{41}}{4} \\ &\approx 1.60 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_4}, \vec{W}_{v_6}) &= \sqrt{\left(\frac{4}{5}\right)^2 + (1)^2 + (1)^2} \\ &= \sqrt{\frac{66}{25}} \\ &= \frac{\sqrt{66}}{5} \\ &\approx 1.62 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_5}, \vec{W}_{v_6}) =& \sqrt{(1)^2 + \left(\tfrac{2}{3}\right)^2 + (1)^2} \\ =& \sqrt{\tfrac{22}{9}} \\ =& \tfrac{\sqrt{22}}{3} \\ \approx& \ 1.56 \end{aligned}$$

Similarly, for L_1 Norm we have

$$d(\vec{W}_{v_i}, \vec{W}_{v_j}) = \sum_{s=1}^{k} |d(N_s(v_i), N_s(v_j))|$$

and for L_∞ Norm we have

$$d(\vec{W}_{v_i}, \vec{W}_{v_j}) = \max_{s \in [k]} |d(N_s(v_i), N_s(v_j))|.$$

Again we skip the details but list our findings in Table 5.3. It appears that using set-valued vectors to represent vertices yields even better results, as different pairs of vertices almost always receive different distances.

Table 5.3: Distance Metrics between each $\vec{W}_{v_i}$'s.

	$L_2\ Norm$	$L_1\ Norm$	$L_\infty\ Norm$
$d(\vec{W}_{v_1}, \vec{W}_{v_2})$	$\frac{2\sqrt{61}}{15}$	$\frac{22}{15}$	$\frac{4}{5}$
$d(\vec{W}_{v_1}, \vec{W}_{v_3})$	$\frac{3}{2}$	$\frac{5}{2}$	1
$d(\vec{W}_{v_1}, \vec{W}_{v_4})$	$\frac{\sqrt{61}}{6}$	$\frac{11}{6}$	1
$d(\vec{W}_{v_1}, \vec{W}_{v_5})$	$\sqrt{3}$	3	1
$d(\vec{W}_{v_1}, \vec{W}_{v_6})$	$\frac{\sqrt{41}}{4}$	$\frac{11}{4}$	1
$d(\vec{W}_{v_2}, \vec{W}_{v_3})$	$\frac{\sqrt{66}}{5}$	$\frac{14}{5}$	1
$d(\vec{W}_{v_2}, \vec{W}_{v_4})$	$\frac{\sqrt{61}}{10}$	$\frac{11}{10}$	$\frac{3}{5}$
$d(\vec{W}_{v_2}, \vec{W}_{v_5})$	$\frac{\sqrt{22}}{3}$	$\frac{8}{3}$	1
$d(\vec{W}_{v_2}, \vec{W}_{v_6})$	$\frac{\sqrt{19}}{3}$	$\frac{7}{3}$	1
$d(\vec{W}_{v_3}, \vec{W}_{v_4})$	$\frac{\sqrt{59}}{5}$	$\frac{13}{5}$	1
$d(\vec{W}_{v_3}, \vec{W}_{v_5})$	$\sqrt{3}$	3	1
$d(\vec{W}_{v_3}, \vec{W}_{v_6})$	$\frac{\sqrt{13}}{4}$	$\frac{5}{4}$	$\frac{3}{4}$
$d(\vec{W}_{v_4}, \vec{W}_{v_5})$	$\frac{\sqrt{41}}{4}$	$\frac{11}{4}$	1
$d(\vec{W}_{v_4}, \vec{W}_{v_5})$	$\frac{\sqrt{66}}{5}$	$\frac{14}{5}$	1
$d(\vec{W}_{v_5}, \vec{W}_{v_6})$	$\frac{\sqrt{22}}{3}$	$\frac{8}{3}$	1

CHAPTER 6

CONCLUDING REMARKS

In this thesis we first surveyed the concept of clustering and known clustering algorithms. We used the phylogenetic tree and the clustering of species as an example. In this example we analyzed the differences between the different clustering algorithms while applying each algorithm to a phylogentic tree.

Next we moved onto exploring the important concepts and techniques in graph clustering where one of the main topics is to distinguish vertices based on the information from the graph structure. Based off the location of each vertex and those adjacent to one another, we will place those in their own cluster. This process relies on the definition and application of the vertex similarities and distance measurements.

For each pair of vertices, they are treated as data points that contain information from the graph structure. With this in mind, we then measure the level of similarities between each pair of vertices by using three of the main distance metrics and set differences.

The three main distance metrics are: The L_2 Norm, L_1 Norm, and L_∞ Norm. We then proceeded onto applying these methods to our two examples, where the results from each method are compared and analyzed. In particular, preliminary study seems to suggest that our proposed model performs well with set distance and the L_2 norm.

As for future work we plan to use our proposed model on much larger networks or graphs of practical interests.

REFERENCES

[1] IBM SPSS Statistics guides: Chapter 16-Cluster Analysis.
http://www.norusis.com/

[2] Biology Stockexchange.
https://biology.stackexchange.com/questions/17185/draw-simplified-phylogenetic-tree-for-kids

[3] Chire. Density-Based Clustering. Clustering Analysis.
https://commons.wikimedia.org/wiki/File:OPTICS-Gaussian-data.svg

[4] A. Condon, R.M. Karp, Algorithms for graph partitioning on the planted partition model, Random Structures & Algorithms 18 (2) (2001) 116–140

[5] M. Cowlishaw, N. Fillmore. Linear Algebra. Introduction to Numerical Analysis.
http://pages.cs.wisc.edu/ amos/412/lecture-notes/lecture14.pdf

[6] P. Erdős, A. Rényi, On random graphs I, in: Selected Papers of Alfréd , vol. 2, Akadémiai Kiadó, Budapest, Hungary, 1976, PP. 308-315. First publication in Publ. Math. Debrecen 1959.

[7] P. Erdős, A. Rényi, On the evolution of random graphs, in" Selected Papers of Alfréd Rényi, vol. , Akadémiai Kiadó, Budapest, Hungary, 1976, pp. 482-525. First publication in MTA Mat. Kut. Int. Kőzl. 1960.

[8] E.N. Gilbert, Random graphs, Annals of Mathematical Statistics 30 (4) (1959) 1141–1144.

[9] Introduction to Probability, Statistics and Random Processes. 1.2.2 Set Operations.
https://www.probabilitycourse.com/chapter1/1_2_2_set_operations.php

[10] Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multi-species coalescent model, PNAS September 11, 2012, vol. 109 no. 37 14942-14947, doi:10.1073/pnas.1211733109
https://phys.org/news/2012-10-forensic-speciation-splicing-genetic-phylogenic.html#jCp

[11] S.E. Schaeffer, Graph Clustering, *Computer Science Review* **1** 2007, 27–64.

[12] S.E. Schaeffer, Stochastic local clustering, in T.B. Ho, D. Cheung , H. Liu (Eds.), Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD, in Lecture Notes in Computer Science, vol. 3518, Springer-Velgar GmbH, Berlin, Heidelberg, Germany, 2005.

[13] P. Shankar. What are applications of combinatorics in computer science? *https://www.quora.com/What-are-applications-of-combinatorics-in-computer-science*

[14] Symmetric Difference. Brilliant.org. *https://brilliant.org/wiki/sets-symmetric-difference/*

[15] S.E. Virtanen, Properties of nonuniform random graph models, Tech. Rep. HUT-TCS-A77, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espo, Finland, May 2003.

[16] E.W. Weisstein, “Caveman Graph”. From MathWorld–A Wolfram Web Resource. *http://mathworld.wolfram.com/CavemanGraph.html*

Appendix A

L_1 NORM AND L_∞ NORM FOR R_{V_I}

Recall the formula:

$$d(\vec{R}_{v_i}, \vec{R}_{v_j}) = \sum_{s=1}^{k} |\delta_s(v_i) - \delta_s(v_j)|.$$

Computation details:

$(\vec{R}_{v_1} - \vec{R}_{v_2}) = |(3-3)| + |(2-2)| + |(0-0)| = 0$

$(\vec{R}_{v_1} - \vec{R}_{v_3}) = |(3-3)| + |(2-1)| + |(0-1)| = 1+1 = 2$

$(\vec{R}_{v_1} - \vec{R}_{v_4}) = |(3-4)| + |(2-1)| + |(0-0)| = 1+1 = 2$

$(\vec{R}_{v_1} - \vec{R}_{v_5}) = |(3-1)| + |(2-2)| + |(0-2)| = 2+2 = 4$

$(\vec{R}_{v_1} - \vec{R}_{v_6}) = |(3-2)| + |(2-2)| + (0-1)| = 1+1 = 2$

$(\vec{R}_{v_2} - \vec{R}_{v_3}) = |(3-3)| + |(2-1)| + |(0-1)| = 1+1 = 2$

$(\vec{R}_{v_2} - \vec{R}_{v_4}) = |(3-4)| + |(2-1)| + |(0-0)| = 1+1 = 2$

$(\vec{R}_{v_2} - \vec{R}_{v_5}) = |(3-1)| + |(2-2)| + |(0-2)| = 2+2 = 4$

$(\vec{R}_{v_2} - \vec{R}_{v_6}) = |(3-2)| + |(2-2)| + |(0-1)| = 1+1 = 2$

$(\vec{R}_{v_3} - \vec{R}_{v_4}) = |(3-4)| + |(1-1)| + |(1-0)| = 1+1 = 2$

$(\vec{R}_{v_3} - \vec{R}_{v_5}) = |(3-1)| + |(1-2)| + |(1-2)| = 2+1+1 = 4$

$(\vec{R}_{v_3} - \vec{R}_{v_6}) = |(3-2)| + |(1-2)| + |(1-1)| = 1+1 = 2$

$(\vec{R}_{v_4} - \vec{R}_{v_5}) = |(4-1)| + |(1-2)| + |(0-2)| = 3+1+2 = 6$

$(\vec{R}_{v_4} - \vec{R}_{v_6}) = |(4-2)| + |(1-2)| + |(0-1)| = 2+1+1 = 4$

$(\vec{R}_{v_5} - \vec{R}_{v_6}) = |(1-2)| + |(2-2)| + |(2-1)| = 1+1 = 2$

Similarly, recall:

$$d(\vec{R}_{v_i}, \vec{R}_{v_j}) = \max_{s \in [k]} |\delta_s(v_i) - \delta_s(v_j)|.$$

Computation details:

$(\vec{R}_{v_1} - \vec{R}_{v_2}) = \max_{s\in[1,6]}(0, 0, 0) = 0$

$(\vec{R}_{v_1} - \vec{R}_{v_3}) = \max_{s\in[1,6]}(0, 1, 1) = 1$

$(\vec{R}_{v_1} - \vec{R}_{v_4}) = \max_{s\in[1,6]}(1, 1, 0) = 1$

$(\vec{R}_{v_1} - \vec{R}_{v_5}) = \max_{s\in[1,6]}(2, 0, 2) = 2$

$(\vec{R}_{v_1} - \vec{R}_{v_6}) = \max_{s\in[1,6]}(1, 0, 1) = 1$

$(\vec{R}_{v_2} - \vec{R}_{v_3}) = \max_{s\in[1,6]}(0, 1, 1) = 1$

$(\vec{R}_{v_2} - \vec{R}_{v_4}) = \max_{s\in[1,6]}(1, 1, 0) = 1$

$(\vec{R}_{v_2} - \vec{R}_{v_5}) = \max_{s\in[1,6]}(2, 0, 2) = 2$

$(\vec{R}_{v_2} - \vec{R}_{v_6}) = \max_{s\in[1,6]}(1, 0, 1) = 1$

$(\vec{R}_{v_3} - \vec{R}_{v_4}) = \max_{s\in[1,6]}(1, 0, 1) = 1$

$(\vec{R}_{v_3} - \vec{R}_{v_5}) = \max_{s\in[1,6]}(2, 1, 1) = 2$

$(\vec{R}_{v_3} - \vec{R}_{v_6}) = \max_{s\in[1,6]}(1, 1, 0) = 1$

$(\vec{R}_{v_4} - \vec{R}_{v_5}) = \max_{s\in[1,6]}(3, 1, 2) = 3$

$(\vec{R}_{v_4} - \vec{R}_{v_6}) = \max_{s\in[1,6]}(2, 1, 1) = 2$

$(\vec{R}_{v_5} - \vec{R}_{v_6}) = \max_{s\in[1,6]}(1, 0, 1) = 1$

Appendix B

L_1 NORM AND L_∞ NORM FOR W_{V_I}

Recall the formula:

$$d(\vec{W}_{v_i}, \vec{W}_{v_j}) = \sum_{s=1}^{k} |d(N_s(v_i), N_s(v_j))|$$

Computational details:

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_2}) &= \frac{4}{5} + \frac{2}{3} + 0 \\ &= \frac{22}{15} \\ &\approx 1.47 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_3}) &= \frac{1}{2} + 1 + 1 \\ &= \frac{5}{2} \\ &\approx 2.5 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_4}) &= \frac{5}{6} + 1 + 0 \\ &= \frac{11}{6} \\ &\approx 1.83 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_5}) &= 1 + 1 + 1 \\ &= 3 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_1}, \vec{W}_{v_6}) &= \frac{3}{4} + 1 + 1 \\ &= \frac{11}{4} \\ &\approx 2.75 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_2}, \vec{W}_{v_3}) &= \frac{4}{5} + 1 + 1 \\ &= \frac{14}{5} \\ &\approx 2.8 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_2}, \vec{W}_{v_4}) &= \frac{3}{5} + \frac{1}{2} + 0 \\ &= \frac{11}{10} \\ &\approx 1.1 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_2}, \vec{W}_{v_5}) &= \frac{2}{3} + 1 + 1 \\ &= \frac{8}{3} \\ &\approx 2.67 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_2}, \vec{W}_{v_6}) &= \frac{1}{3} \ + \ 1 \ + \ 1 \\ &= \frac{7}{3} \\ &\approx 2.33 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_3}, \vec{W}_{v_4}) &= \frac{3}{5} \ + \ 1 \ + \ 1 \\ &= \frac{13}{5} \\ &\approx 2.6 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_3}, \vec{W}_{v_5}) &= 1 \ + \ 1 \ + \ 1 \\ &= 3 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_3}, \vec{W}_{v_6}) &= \frac{3}{4} \ + \ \frac{1}{2} \ + \ 0 \\ &= \frac{5}{4} \\ &\approx 1.25 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_4}, \vec{W}_{v_5}) &= \frac{3}{4} \ + \ 1 \ + \ 1 \\ &= \frac{11}{4} \\ &\approx 2.75 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_4}, \vec{W}_{v_6}) &= \frac{4}{5} \ + \ 1 \ + \ 1 \\ &= \frac{14}{5} \\ &\approx 2.8 \end{aligned}$$

$$\begin{aligned} d(\vec{W}_{v_5}, \vec{W}_{v_6}) &= 1 \ + \ \frac{2}{3} \ + \ 1 \\ &= \frac{8}{3} \\ &\approx 2.67 \end{aligned}$$

Similarly, recall:

$$d(\vec{W}_{v_i}, \vec{W}_{v_j}) = \max_{s\in[k]} |d(N_s(v_i), N_s(v_j))|$$

Computation details:

$$d(\vec{W}_{v_1}, \vec{W}_{v_2}) = \max_{s\in[k]} |d(N_i(v_1), N_s(v_2))| = \max_{1\in[6]}(\frac{4}{5}, \frac{2}{3}, 0) = \frac{4}{5}$$

$$d(\vec{W}_{v_1}, \vec{W}_{v_3}) = \max_{s\in[k]} |d(N_s(v_1), N_s(v_3))| = \max_{1\in[6]}(\frac{1}{2}, 1, 1) = 1$$

$$d(\vec{W}_{v_1}, \vec{W}_{v_4}) = \max_{s\in[k]} |d(N_s(v_1), N_s(v_4))| = \max_{1\in[6]}(\frac{5}{6}, 1, 0) = 1$$

$$d(\vec{W}_{v_1}, \vec{W}_{v_5}) = \max_{s\in[k]} |d(N_s(v_1), N_s(v_5))| = \max_{1\in[6]}(1, 1, 1) = 1$$

$$d(\vec{W}_{v_1}, \vec{W}_{v_6}) = \max_{s\in[k]} |d(N_s(v_1), N_s(v_6))| = \max_{1\in[6]}(\frac{3}{4}, 1, 1) = 1$$

$$d(\vec{W}_{v_2}, \vec{W}_{v_3}) = \max_{s\in[k]} |d(N_s(v_2), N_s(v_3))| = \max_{1\in[6]}(\frac{4}{5}, 1, 1) = 1$$

$$d(\vec{W}_{v_2}, \vec{W}_{v_4}) = \max_{s\in[k]} |d(N_s(v_2), N_s(v_4))| = \max_{1\in[6]}(\frac{3}{5}, \frac{1}{2}, 0) = \frac{3}{5}$$

$$d(\vec{W}_{v_2}, \vec{W}_{v_5}) = \max_{s\in[k]} |d(N_s(v_2), N_s(v_5))| = \max_{1\in[6]}(\frac{2}{3}, 1, 1) = 1$$

$$d(\vec{W}_{v_2}, \vec{W}_{v_6}) = \max_{s\in[k]} |d(N_s(v_2), N_s(v_6))| = \max_{1\in[6]}(\frac{1}{3}, 1, 1) = 1$$

$$d(\vec{W}_{v_3}, \vec{W}_{v_4}) = \max_{s\in[k]} |d(N_s(v_3), N_s(v_4))| = \max_{1\in[6]}(\frac{3}{5}, 1, 1) = 1$$

$$d(\vec{W}_{v_3}, \vec{W}_{v_5}) = \max_{s\in[k]} |d(N_s(v_3), N_s(v_5))| = \max_{1\in[6]}(1, 1, 1) = 1$$

$$d(\vec{W}_{v_3}, \vec{W}_{v_6}) = \max_{s\in[k]} |d(N_s(v_3), N_s(v_6))| = \max_{1\in[6]}(\frac{3}{4}, \frac{1}{2}, 0) = \frac{3}{4}$$

$$d(\vec{W}_{v_4}, \vec{W}_{v_5}) = \max_{s\in[k]} |d(N_s(v_4), N_s(v_5))| = \max_{1\in[6]}(\frac{3}{4}, 1, 1) = 1$$

$$d(\vec{W}_{v_4}, \vec{W}_{v_6}) = \max_{s\in[k]} |d(N_s(v_4), N_s(v_6))| = \max_{1\in[6]}(\frac{4}{5}, 1, 1) = 1$$

$$d(\vec{W}_{v_5}, \vec{W}_{v_6}) = \max_{s\in[k]} |d(N_s(v_5), N_s(v_6))| = \max_{1\in[6]}(1, \frac{2}{3}, 1) = 1$$

Appendix C

VERIFICATION OF THE TRIANGLE INEQUALITY

We are interested in showing

$$d(\vec{W}_{v_i}, \vec{W}_{v_j}) \ \leq \ d(\vec{W}_{v_i}, \vec{W}_{v_k}) + d(\vec{W}_{v_k}, \vec{W}_{v_j})$$

of our defined distance measure with the example used.

- $d(\vec{W}_{v_1}, \vec{W}_{v_3}) \ \leq \ d(\vec{W}_{v_1}, \vec{W}_{v_2}) + d(\vec{W}_{v_2}, \vec{W}_{v_3})$

  $\left(\frac{1}{2}, 1, 1\right) \leq \left(\frac{4}{5}, \frac{2}{3}, 0\right) + \left(\frac{4}{5}, 1, 1\right)$

  $\left(\frac{1}{2}, 1, 1\right) \leq \left(\frac{8}{5}, \frac{5}{3}, 1\right) \checkmark$

- $d(\vec{W}_{v_1}, \vec{W}_{v_4}) \ \leq \ d(\vec{W}_{v_1}, \vec{W}_{v_3}) + d(\vec{W}_{v_3}, \vec{W}_{v_4}) \quad or \quad d(\vec{W}_{v_1}, \vec{W}_{v_4}) \ \leq \ d(\vec{W}_{v_1}, \vec{W}_{v_2}) + d(\vec{W}_{v_2}, \vec{W}_{v_4})$

  $\left(\frac{5}{6}, 1, 0\right) \leq \left(\frac{1}{2}, 1, 1\right) + \left(\frac{3}{5}, 1, 1\right) \quad or \quad \left(\frac{5}{6}, 1, 0\right) \leq \left(\frac{4}{5}, \frac{2}{3}, 0\right) + \left(\frac{3}{5}, \frac{1}{2}, 0\right)$

  $\left(\frac{5}{6}, 1, 0\right) \leq \left(\frac{11}{10}, 2, 2\right) \checkmark \quad or \quad \left(\frac{5}{6}, 1, 0\right) \leq \left(\frac{7}{5}, \frac{7}{6}, 0\right) \checkmark$

- $d(\vec{W}_{v_1}, \vec{W}_{v_5}) \ \leq \ d(\vec{W}_{v_1}, \vec{W}_{v_2}) + d(\vec{W}_{v_2}, \vec{W}_{v_5}) \quad or \quad d(\vec{W}_{v_1}, \vec{W}_{v_5}) \ \leq \ d(\vec{W}_{v_1}, \vec{W}_{v_3}) + d(\vec{W}_{v_3}, \vec{W}_{v_5})$

  $(1, 1, 1) \leq \left(\frac{1}{2}, 1, 1\right) + \left(\frac{3}{5}, 1, 1\right) \quad or \quad (1, 1, 1) \leq \left(\frac{4}{5}, \frac{2}{3}, 0\right) + \left(\frac{3}{5}, \frac{1}{2}, 0\right)$

  $(1, 1, 1) \leq \left(\frac{11}{10}, 2, 2\right) \checkmark \quad or \quad (1, 1, 1) \leq \left(\frac{7}{5}, \frac{7}{6}, 0\right) \checkmark$

  $d(\vec{W}_{v_1}, \vec{W}_{v_5}) \ \leq \ d(\vec{W}_{v_1}, \vec{W}_{v_4}) + d(\vec{W}_{v_4}, \vec{W}_{v_5})$

  $(1, 1, 1) \leq \left(\frac{5}{6}, 1, 0\right) + \left(\frac{3}{4}, 1, 1\right)$

  $(1, 1, 1) \leq \left(\frac{19}{12}, 2, 1\right) \checkmark$

- $$\begin{array}{lcl} d(\vec{W}_{v_1},\vec{W}_{v_6}) \ \le \ d(\vec{W}_{v_1},\vec{W}_{v_2}) + d(\vec{W}_{v_2},\vec{W}_{v_6}) & or & d(\vec{W}_{v_1},\vec{W}_{v_6}) \ \le \ d(\vec{W}_{v_1},\vec{W}_{v_3}) + d(\vec{W}_{v_3},\vec{W}_{v_6}) \\ \left(\frac{3}{4},1,1\right) \le \left(\frac{4}{5},\frac{2}{3},0\right) + \left(\frac{1}{3},1,1\right) & or & \left(\frac{3}{4},1,1\right) \le \left(\frac{1}{2},1,1\right) + \left(\frac{3}{4},\frac{1}{2},0\right) \\ \left(\frac{3}{4},1,1\right) \le \left(\frac{17}{15},\frac{5}{3},1\right) \checkmark & or & \left(\frac{3}{4},1,1\right) \le \left(\frac{5}{4},\frac{3}{2},1\right) \checkmark \end{array}$$

$$\begin{array}{lcl} d(\vec{W}_{v_1},\vec{W}_{v_6}) \ \le \ d(\vec{W}_{v_1},\vec{W}_{v_4}) + d(\vec{W}_{v_4},\vec{W}_{v_6}) & or & d(\vec{W}_{v_1},\vec{W}_{v_6}) \ \le \ d(\vec{W}_{v_1},\vec{W}_{v_5}) + d(\vec{W}_{v_5},\vec{W}_{v_6}) \\ \left(\frac{3}{4},1,1\right) \le \left(\frac{5}{6},1,0\right) + \left(\frac{4}{5},1,1\right) & or & \left(\frac{3}{4},1,1\right) \le (1,1,1) + \left(1,\frac{2}{3},1\right) \\ \left(\frac{3}{4},1,1\right) \le \left(\frac{49}{30},2,1\right) \checkmark & or & \left(\frac{3}{4},1,1\right) \le \left(2,\frac{5}{3},2\right) \checkmark \end{array}$$

- $$\begin{array}{l} d(\vec{W}_{v_2},\vec{W}_{v_4}) \ \le \ d(\vec{W}_{v_2},\vec{W}_{v_3}) + d(\vec{W}_{v_3},\vec{W}_{v_4}) \\ \left(\frac{3}{5},\frac{1}{2},0\right) \le \left(\frac{4}{5},1,1\right) + \left(\frac{3}{5},1,1\right) \\ \left(\frac{3}{5},\frac{1}{2},0\right) \le \left(\frac{7}{5},2,2\right) \checkmark \end{array}$$

- $$\begin{array}{lcl} d(\vec{W}_{v_2},\vec{W}_{v_5}) \ \le \ d(\vec{W}_{v_2},\vec{W}_{v_3}) + d(\vec{W}_{v_3},\vec{W}_{v_5}) & or & d(\vec{W}_{v_2},\vec{W}_{v_5}) \ \le \ d(\vec{W}_{v_2},\vec{W}_{v_4}) + d(\vec{W}_{v_4},\vec{W}_{v_5}) \\ \left(\frac{2}{3},1,1\right) \le \left(\frac{4}{5},1,1\right) + (1,1,1) & or & \left(\frac{2}{3},1,1\right) \le \left(\frac{3}{5},\frac{1}{2},0\right) + \left(\frac{3}{4},1,1\right) \\ \left(\frac{2}{3},1,1\right) \le \left(\frac{9}{5},2,2\right) \checkmark & or & \left(\frac{2}{3},1,1\right) \le \left(\frac{27}{20},\frac{3}{2},1\right) \checkmark \end{array}$$

- $$\begin{array}{lcl} d(\vec{W}_{v_2},\vec{W}_{v_6}) \ \le \ d(\vec{W}_{v_2},\vec{W}_{v_3}) + d(\vec{W}_{v_3},\vec{W}_{v_6}) & or & d(\vec{W}_{v_2},\vec{W}_{v_6}) \ \le \ d(\vec{W}_{v_2},\vec{W}_{v_4}) + d(\vec{W}_{v_4},\vec{W}_{v_6}) \\ \left(\frac{1}{3},1,1\right) \le \left(\frac{4}{5},1,1\right) + \left(\frac{3}{4},\frac{1}{2},0\right) & or & \left(\frac{1}{3},1,1\right) \le \left(\frac{3}{5},\frac{1}{2},0\right) + \left(\frac{4}{5},1,1\right) \\ \left(\frac{1}{3},1,1\right) \le \left(\frac{31}{20},\frac{3}{2},1\right) \checkmark & or & \left(\frac{1}{3},1,1\right) \le \left(\frac{7}{5},\frac{3}{2},1\right) \checkmark \end{array}$$

$$\begin{aligned} d(\vec{W}_{v_2},\vec{W}_{v_6}) \ &\le \ d(\vec{W}_{v_2},\vec{W}_{v_5}) + d(\vec{W}_{v_5},\vec{W}_{v_6}) \\ \left(\frac{1}{3},1,1\right) &\le \left(\frac{2}{3},1,1\right) + \left(1,\frac{2}{3},1\right) \\ \left(\frac{1}{3},1,1\right) &\le \left(\frac{5}{3},\frac{5}{3},2\right) \checkmark \end{aligned}$$

- $$\begin{aligned} d(\vec{W}_{v_3},\vec{W}_{v_5}) \ &\le \ d(\vec{W}_{v_3},\vec{W}_{v_4}) + d(\vec{W}_{v_4},\vec{W}_{v_5}) \\ (1,1,1) &\le \left(\frac{3}{5},1,1\right) + \left(\frac{3}{4},1,1\right) \\ (1,1,1) &\le \left(\frac{27}{20},2,2\right) \checkmark \end{aligned}$$

- $$\begin{aligned} d(\vec{W}_{v_3},\vec{W}_{v_6}) \ &\le \ d(\vec{W}_{v_3},\vec{W}_{v_4}) + d(\vec{W}_{v_4},\vec{W}_{v_6}) & \quad \textit{or} \quad & d(\vec{W}_{v_3},\vec{W}_{v_6}) \ \le \ d(\vec{W}_{v_3},\vec{W}_{v_5}) + d(\vec{W}_{v_5},\vec{W}_{v_6}) \\ \left(\frac{3}{4},\frac{1}{2},0\right) &\le \left(\frac{3}{5},1,1\right) + \left(\frac{4}{5},1,1\right) & \quad \textit{or} \quad & \left(\frac{3}{4},\frac{1}{2},0\right) \le (1,1,1) + \left(1,\frac{2}{3},1\right) \\ \left(\frac{3}{4},\frac{1}{2},0\right) &\le \left(\frac{7}{5},2,2\right) \checkmark & \quad \textit{or} \quad & \left(\frac{3}{4},\frac{1}{2},0\right) \le \left(2,\frac{5}{3},2\right) \checkmark \end{aligned}$$

- $$\begin{aligned} d(\vec{W}_{v_4},\vec{W}_{v_6}) \ &\le \ d(\vec{W}_{v_4},\vec{W}_{v_5}) + d(\vec{W}_{v_5},\vec{W}_{v_6}) \\ \left(\frac{4}{5},1,1\right) &\le \left(\frac{3}{4},1,1\right) + \left(1,\frac{2}{3},1\right) \\ \left(\frac{4}{5},1,1\right) &\le \left(\frac{7}{4},\frac{5}{3},2\right) \checkmark \end{aligned}$$