SIMC2.0 Junior Challenge 2022 REPORT: Pandemics over networks

Commonwealth Secondary School (Michelle Aurelia Yudianto, Aarav Malik, Pua Hong Wei)

<u>Overview</u>

In this problem, we investigate the infection of people by Duovid in the village. A person may only be infected if there are at least 2 people they are in contact with.

1.1 Spreading on lattices

Denote that we use the colour red, to represent infected nodes and black to represent uninfected nodes

Triangular Village

In triangular Village, we use 'upright triangles' to refer to a unit triangle where an edge is at the bottom and 'reversed triangles' is a unit triangle with an edge at the top. Lastly, we define the nodes by numbers, going from the top row to the bottom row, then left to right, as shown in *Figure 1.0.*

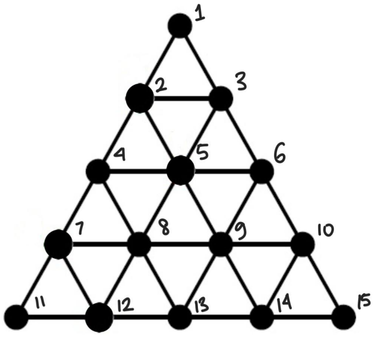


Figure 1.0

Question 1(a): 2

We prove that there is no smaller number n_0 than 2. Since a person needs to be in touch with at least 2 infected people to get infected, the virus can only spread if 2 or more people are infected initially. We show that when n_0 is 2, it is possible for the virus to spread to the whole Triangular Village. Let B and C be the initial infected people, then everyone will be infected at the end. Hence, the minimal number n_0 is **2**.

Question 1(b): $\frac{22}{35}$

Using n_0= 2 from Question 1(a), we will show that as long as the 2 initial infected nodes will cause 2 nodes connected by an edge to be infected, then the entire village will be infected. We will prove this using Lemma 1.

Lemma 1: If an infected node is 2 or fewer edges away from another infected node, then the entire village will be infected.

Proof: When the nodes are 1 edge away (adjacent), there is always a node, let it be a_1, that is connected to the initial 2 adjacent nodes. Thus, a_1 will always be infected and will form a triangle consisting of three adjacent infected nodes as shown in *Figure 1.1.* At least one of the edges of the triangle will always be a part of another triangle, and therefore create another infected triangle like *Figure 1.1.* This process will continue until eventually the entire village is infected. When the nodes are 2 edges away (Like *Figure 1.3* and *Figure 1.4)*, there will always be one node which is adjacent to those 2 nodes (hence them being 2 edges away- they have to connect through a node). That will cause that particular node to get infected and then there will be 2 adjacent nodes, which as stated earlier will infect the rest of the village.

Using Lemma 1, for the entire village to be infected, we need 2 nodes which are 2 or fewer edges away from each other to be infected. There are 3 cases where this happens, shown in *Figure 1.2*, *Figure 1.3* and

Figure 1.4. We need to calculate the total combinations of 2 nodes being chosen from 15 nodes. This is ${}_{15}C_{2}$ $= \frac{15!}{2!\times13!} = 105$

Followed by that, we need to calculate the total number of combinations in *Figure 1.2*, *Figure 1.3* and *Figure 1.4*. The total number of combinations for *Figure 1.2* is the total number of edges. Upon further inspection, this is the 4th triangular number multiplied by 3 (there are the 4th triangular number of upright triangles, in which every edge is accounted for and no edge is accounted for twice). This results in $3(1 + 2 + 3 + 4) = 3 \times \frac{4(4+1)}{2} = 30$. The total number of combinations in *Figure 1.3* is the number of reversed triangles (the 3rd triangle number) multiplied by 3 (As shown in *Figure 1.3*, one of these patterns is made up of an upright and a reversed triangle, so the number of cases for this is whichever is less of the upright triangles and reversed triangles, where the reversed triangles are less. The multiplication by 3 is to account for rotation) which is $3((1 + 2 + 3 + 4) - 4) = 3(1 + 2 + 3) = 3 \times \frac{3(3+1)}{2} = 18$. In *Figure 1.4*, the number of cases would be the number of 3 in a row, which would be the number of upright triangles minus the last row (The 4th triangular number - 4, which is the 3rd triangular number) (as the number of 3 in a row is the same as the number of upright triangles minus 1 for every row) multiplied by 3 (to account for rotation). This would be $3((1 + 2 + 3 + 4) - 4) = 3(1 + 2 + 3)$ $= 3 \times \frac{3(3+1)}{2} = 18$. Thus, the chance that the entire village will be infected is $\frac{30+18+18}{105} = \frac{66}{105} = \frac{22}{35}$

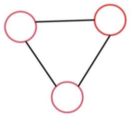


Figure 1.1

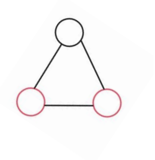


Figure 1.2

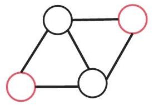


Figure 1.3

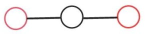


Figure 1.4

Question 1(c): 3.

Using Lemma 1, for the entire village to be infected, we need 2 nodes which are 2 or fewer edges away from each other to be infected. Therefore, for the village to not be infected with rotations and reflections accounted for, there are 4 different types of points. The tip, tip-adjacent, centre-side and centre nodes. The tip nodes are 1, 11 and 15, the tip-adjacent nodes are 2, 3, 7, 10, 12 and 14, the centre-side 4, 6 and 13 and the centres 5, 8 and 9. Shown in *Figure 1.5*, *Figure 1.6*, *Figure 1.7* and *Figure 1.8* are the nodes 1 or 2 edges away from the circled node (showing tip, tip-adjacent, centre-side and centre nodes respectively). For tips, tip-adjacents and centres, the space remaining can be used for at most 2 nodes. For centre-sides, the space remaining can be used for at most 1 node. When putting a 4th infected node on Triangle Village, one of the 3 cases in *Figure 1.2*, *Figure 1.3* or *Figure 1.4* will happen, thus infecting the entire village by Lemma 1. Therefore, the answer is 1+2=3.

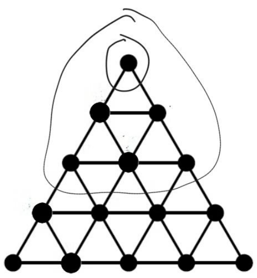


Figure 1.5

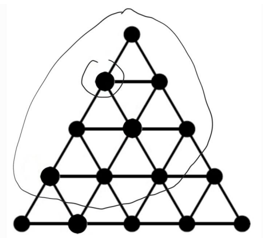


Figure 1.6

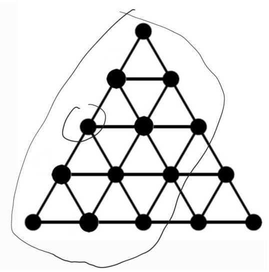


Figure 1.7

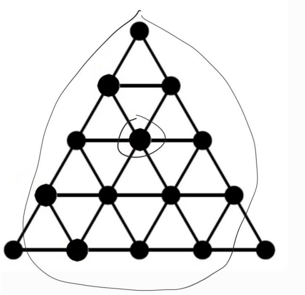


Figure 1.8

Question 1(d): 2, $\frac{36L-12}{L^3+6L^2+11L+6}$

When L can take on any positive integer, we can prove using Lemma 1, that as long as 2 infected nodes are 2 or fewer edges away, the entire village will get infected. Obviously, when n_0 is less than 2, the virus cannot spread and therefore the entire village will never get infected. Therefore, the minimal number n_0 of initially infected people when L can take on any positive integer is 2.

Similar to question 1(b), there are 3 scenarios of the initial 2 infected people for the virus to spread to the rest of Triangular Village as shown in *Figure 1.2, Figure 1.3*, and *Figure 1.4.* Again, there are 3 different rotations of each scenario, and for simplicity, let us only calculate the number of combinations of each rotation and multiply it by 3. Let the nth row in the triangular village be counted from top to bottom. Let us use the orientation as shown in *Figure 1.2, Figure 1.3*, and *Figure 1.4.* For the scenario shown in *Figure 1.2,* there is 1 combination when it is located in the 1st row, 2 combinations in the 2nd row, 3 combinations in the 3rd row, and eventually L combinations in the Lth row. Consequently, there will be $3(1 + 2 + 3 +\ldots + L) = 3(\frac{(L)(L+1)}{2})$ combinations for *Figure 1.2*. The number of reversed triangles will be the same as the number of cases for *Figure 1.3* (As shown in *Figure 1.3*, one of these patterns is made up of an upright and a reversed triangle, so the number of cases for this is whichever is less of the upright triangles and reversed triangles, where the reversed triangles are less.), coming to $3((1 + 2 +\ldots + L) - L) = 3((0 + 1 + 2 +\ldots + (L - 1)= 3(\frac{(L-1)(L)}{2})$. In *Figure 1.4*, the number of cases would be the number of 3 in a row, which would be the number of upright triangles minus the last row (The Lth triangular number - L, which is the (L-1)th triangular number) (as the number of 3 in a row is the same as the number of upright triangles minus 1 for every row) multiplied by 3 (to account for rotation). There will be $3((1 + 2 +\ldots + L) - L) = 3((0 + 1 + 2 +\ldots + (L - 1) = 3(\frac{(L-1)(L)}{2})$ combinations for each rotation of *Figure 1.4.* Thus, there are $3(\frac{(L)(L+1)}{2}+\frac{(L-1)(L)}{2}+\frac{(L-1)(L)}{2}) = \frac{9L^2-3L}{2}$ combinations of all the rotations of *Figure 1.2, Figure 1.3,* and *Figure 1.4.*

Now, we need to calculate the number of ways to choose 2 nodes from all the possible nodes. The total number of nodes in the village when the village length is L is $\frac{(L+1)(L+2)}{2}$. The number of ways to choose 2 nodes from all the possible nodes is ${}_{\frac{(L+1)(L+2)}{2}}C_2$ which is equal to $\frac{L(L+1)(L+2)(L+3)}{8}$. Hence, the chance that the whole village will be infected is $\frac{\frac{9L^2-3L}{2}}{\frac{L(L+1)(L+2)(L+3)}{8}} = \frac{36L-12}{(L+1)(L+2)(L+3)} = \frac{36L-12}{L^3+6L^2+11L+6}$.

Question 2(a): Node no 8

The only nodes which would be directly infected from B and C are A and node no 8. Vaccinating A will only reduce the number of infected people by 1, while vaccinating node 8 would only result in A being infected, thus infecting 3 people, which is the minimum. Vaccinating anyone else will result in A, B, C and 8 all being infected, which is 4 people being infected, and is more than the 3 infected people when node no 8 is vaccinated.

Question 2(b): Node no 8 or Node no 9

To reduce calculation, let us divide the combination into different cases of who gets the vaccine where in each case, the node can be obtained by either rotation and/or reflection of the other nodes in the same case. Since nodes 2 and 5 are already infected, we can exclude them from the cases. Let case 1 consist of nodes 1,11,15, case 2 consist of nodes 3,7,10,12,13, case 3 consist of nodes 4,6,13, and case 4 consist of nodes 8 and 9. A person cannot be infected by the virus if and only if it only has at most 1 infected node that the person is in touch with. Therefore, we need to choose the person to vaccinate such that it can block the most number of nodes from having more than 1 infected node that it is in touch with.

In case 1, WLOG node 1 is vaccinated, it won't be able to prevent anyone from being infected because node 3 will still have 3 infected people that it is in touch with. Similarly, if node 11 is vaccinated, nodes 7 and 12 will still have at least 2 more infected nodes that it is in touch with. The same goes for when node 15 is vaccinated. So, there are 14 infected people at the end of case 1.

In case 2, the node can block 1 other node from being infected. WLOG node 3 is vaccinated, node 1 will only have one infected node that it is in touch with. Meanwhile, the other nodes will still have at least 2 infected nodes that they are in touch with. The same applies to nodes 7,10,12 and 14. Hence, for case 2, 13 people will be infected in the end.

In case 3, the vaccinated nodes will not be able to protect any other nodes from getting infected. WLOG, if node 13 is vaccinated, every node in touch with it will still have at least 2 other infected nodes, other than 13, that they are in touch with such that they will still get infected. Hence, there are 14 infected people at the end of case 3.

Finally, in case 4, the vaccinated nodes will be able to protect 3 other nodes from getting infected. WLOG, if node 8 is vaccinated, node 7 is only in touch with 1 infected node, which is node 4, while node 12 is also only in touch with 1 infected node, which is node 13. Then, both of the nodes that node 11 is in touch with are not infected. Thus, nodes 7,8,11,12 will not get infected at the end. Similarly, when node 9 is infected, nodes 9,10,14,15 will also not get infected at the end. Hence, there will be 11 people who are infected for case 4 at the end.

Therefore, we can conclude that either node 8 or node 9 should be given the vaccine such that the number of infected people will be minimised to only 11 people.

Square Village

In Square Village, We number the nodes from top to bottom then left to right. See *Figure 3.0.*

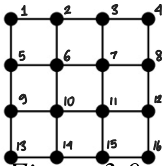


Figure 3.0

Question 3(a): 4
First we show that if n_0= 1, it is impossible for Duvoid to spread to the entire Square Village. When n_0=1, it is obvious that the virus cannot spread because a node requires it to be in contact with 2 other infected nodes for it to be infected, it is clear that this is not possible if there is only 1 initial infected node.

However, n_0= 2 and 3 not infecting the entire square village is a bit harder to prove.

Lemma 2: The maximum number of people that the initial infected nodes can infect is the number of people that is in the smallest rectangle that encloses the initial nodes.

Proof: There are 2 ways for Duovid to spread in the Square Village - by 2 nodes being diagonal and 2 nodes being 2 edges away in a straight line, as shown in *Figures 3.1* and *3.2*. In the first case, the new node will share the same row as one of the previously infected nodes and it will also share the same column as another of the infected nodes. As a node outside the original minimum rectangle has to have at least one of the rows or columns different from nodes inside the rectangle, that particular node will be inside the rectangle. For the 2 edges away in a straight line case, the infected node will be in between the 2 original infected nodes, therefore being inside the rectangle. Hence, the infected nodes can only infect those in the enclosed rectangle and it is the maximum number of nodes that any first initial node can infect.

For the smallest rectangle to be the Square Village, all 4 outside sides must have at least 1 node (otherwise the rectangle will be smaller than 4 nodes x 4 nodes and some nodes will be uninfected). It is worth noting that nodes in the 4 by 4 square can only have 3 numbers of outside sides they can border- 0 (nodes no 6, 7, 10, 11, the centre 4), 2 (nodes no 1, 4, 13, 16, the corners) and 1 (all others (nodes no 2, 3, 5, 8, 9, 12, 14, 15), hence known as edge nodes from now onwards). It is obvious that with these restrictions, n_0= 2 will not work as the only case which fits these restrictions are the 2 opposite corners, which will not infect anyone else.

When n_0= 3 there must be at least 1 corner node to be infected to fulfil the earlier requirements (If all are edge nodes, they only border 3 outside sides). Also, in order for the virus to spread, at least 2 of the 3 nodes must cause another node that is in contact with both of the nodes to be infected. Otherwise, the infection will stop completely. We split this into cases.

Case 1: 2 opposite corners and 1 other
In this case, the other will only be close enough to infect others with at most 1 of the 2 corner nodes (as the 2 opposite corner nodes are 6 edges apart, and thus the other node cannot be 2 or fewer edges apart from both nodes) and it will not be able to infect the entire 4 x 4.

Case 2: 1 corner and 2 edge nodes (on the sides where the corner is not at)
Case 3: 2 adjacent corners and 1 edge node (on the side where the 2 corners are not at)
In these cases, no node will be close enough to infect another node.

Therefore, n_0= 3 will not infect all 16 villagers.

An example for n_0= 4 will be a diagonal from 1 corner to the opposite corner. Therefore n_0= 4 will infect the entire Square Village.

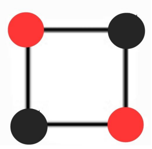


Figure 3.1

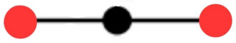


Figure 3.2

Question 3(b):

n_o in the Square Village is different from n_0 in the Triangular Village. While n_0 in the Triangular Village is 2, n_0=2 is not possible in the Square Village, and the minimum n_0 must be 3. This is because while there are 3 different ways for a node to be infected, as shown in *Figures 1.2, 1.3,* and *1.4*, there are only 2 different ways for a node to be infected, as shown in *Figures 3.1* and *3.2*. Consequently, the probability of a node being infected in the Square Village is less than in the Triangular Village. Thus, n_0 is greater in the Square Village compared to the Triangular Village.

For the Square Village to have an equal n_0 as the Triangular Village, we need to create one more edge for each 2 x 2 square (shown in *Figure 3.3*) such that for each node to get infected, there are 3 different ways, similar to in the Triangular Village, as shown in *Figures 1.2, 1.3,* and *1.4*. This will cause the Square Village to have an equal n_0 of value 2 as the one in the Triangular Village. We show that it is possible for everyone in this version of the Square Village to be infected in *Figure 3.4*.

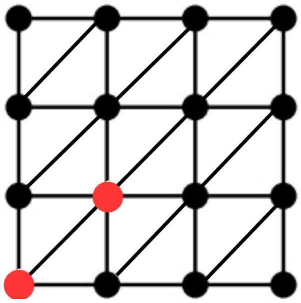


Figure 3.3

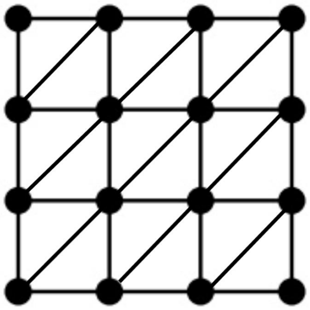
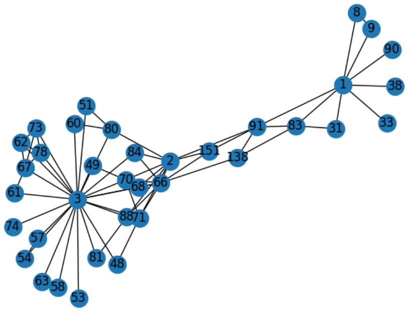


Figure 3.4

1.2 Spreading on complex network (Modelling and Machine Learning)

In this section we used Python Libraries such as NetworkX, sci-kit learn and Matplotlib. With reference to documentation for each stated above along with discussions to debug our code, we managed to accomplish the task. Using NetworkX, we are using nodes to represent people and the edges (connections) between them as points of contact, similar to contact tracing. As for scikit- learn, it is used to code the models required for machine learning. Finally, matplotlib is used to ease the visualisation of graphs and data points.



Question 4(a): 22

Question 4(b): Node 91

Question 5(a):
Firstly, we organised the data in every array to match with the nd, an array of the node labels, in ascending numerical order. This is to ensure accuracy in terms of data plotting.

Next, we had to decide on the vectors to be used, one would be the effectiveness of the vaccination when allocated to each node (x). For the other, we shortlisted vectors with promising results, and after multiple rounds of testing, we determined that the distance between node '151' and each node (path lengths)(y) would be the 2nd vector in the model. The reason for choosing node '151' is that it is the major cause of the spread of the virus since it acts as a bridge through which the virus passes to reach the densely populated area around node '3'. With further observation, we also noticed that the closer the vaccinated node is to 151, the lesser the number of nodes which will be infected at the end. This shows that there may be a relation between the distance of the vaccinated node to node '151' and the effectiveness of the vaccination.

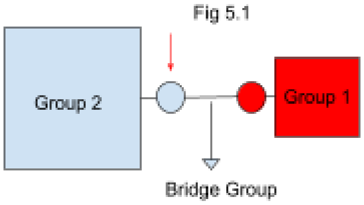
For the prediction purposes, we take in a node input, asking the user to input a node of their choice. We then use this input to derive the index of the node in nd which is then stored in a variable used to locate node distance stored in `plens`, an array of path lengths. Since sci-kit learn does not have a feature to predict x values given the y values, we used the straight line equation, $y = mx + c$, to form an inverse function of the model. By using the following code: `x = (contacts - model.intercept_)/model.coef_` we managed to accomplish this. We obtained the coefficient of x as well as the intercept of the y-axis using in-built functions and then used some basic algebraic manipulation to make x the subject of the equation.

The x value is then printed as output.

Question 5(b):
To draw insights, we used linear regression, a linear approach to modelling to predict vaccination effectiveness. Essentially, we are plotting data points on a linear plane, and then finding the best fit line, $y = mx + c$, to predict future or other possible values using the same straight line.

Since the city has established contact tracing and has information on the contact network of all citizens, we would like to suggest which groups should be vaccinated first, from our current model. After inputting different node numbers into our model, we propose that people who connect groups to other bigger groups, acting as bridges should be focused upon. In the scenario that a group or a person on this bridge is infected, the best option would be to vaccinate the people in the closest contact with them so as to isolate the virus and prevent further spread. To generalise every possible scenario we would be using Fig 5.1. The highlighted red represents the infected areas whilst highlighted blue represents the non-infected ones. As shown in Fig 5.1, the person nearest to the blue Group 2 should be vaccinated so as to isolate the virus entirely to the right side. This is indeed the case in our model as node '151' which is on the bridge between two groups has been infected.

Fig 5.1



The best possible option would be to block off the virus completely in Group 1 itself by vaccinating the person nearest to it if he/she has not been infected yet. To determine which person on the Bridge group to vaccinate, we calculate the length of the shortest path between each person and an infected person in the bridge group.

To determine the bridge group in the city, we simply need to classify people into infected and non-infected and then see which people link these two groups together. These people would most likely be the frontline workers who treat the infected people and return to society to meet their families and continue with their daily activities. Thus, they should be the first group to be vaccinated.

Question 5(c):

In our opinion, simple linear regression may not be the most effective way to cope with the situation in the city as there are other models which would result in higher accuracy and more reliable results.

The first possibility would be using neural networks. We could use a variety of parameters and use them to classify a large number of nodes into neurons. In this case, every node could be run through parameters such as number of degrees and shortest distance from infected nodes, and their location. A simple illustration of the neural network can be seen in Fig 5.2.

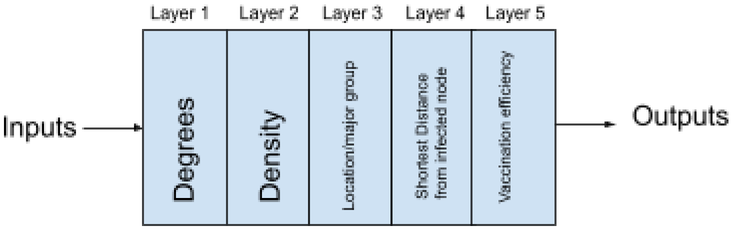


Fig 5.2

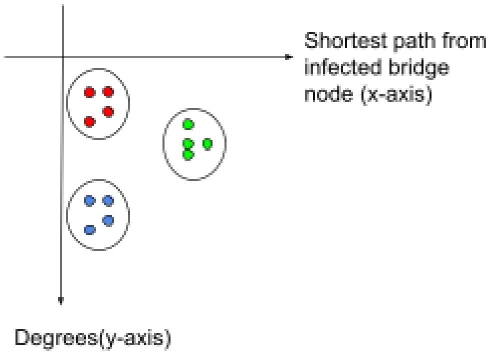
To determine things like location, multiple layers could be used in order to classify it into one of the three major groups: infected, non-infected or bridge. As for the number of degrees, it would determine to some extent, the density of the area around the node which would aid in classification of location. This would further allow in determining and classifying nodes into groups based on the efficiency of giving the vaccine to them. Shortest distance from infected nodes could be used as the final layer before classifying the nodes into vaccination efficiency to ensure higher accuracy.

The outputs could be given in a line graph to ease visualisation. A similar but tweaked approach could be used for tree-based models such as decision trees as they are deterministic unlike neural networks.

A neural network would be very efficient as well because the only required change would be its inputs as the parameters would remain the same for any sample size. Instead of being limited to only two variables in simple linear regression, we can use a large number of parameters in neural networks. Another disadvantage of simple linear regression would be that if there is no common trend or if there are outliers in the data points, the accuracy of the predictions would be decreased significantly as simple linear regression is limited to predicting values using a straight line and one would have to upgrade to multiple linear regression if they want to introduce external variables. Even if after that, you are still predicting points using a straight line and outliers would still significantly pull down the accuracy.

Another possibility would be using Support Vector Machines (SVMs). Using a 3 dimensional graph could improve accuracy but in terms of classification. We would keep slicing and zooming in on certain aspects of the machine to further narrow down the results. A simple illustration of the model can be seen in Fig 5.3 and Fig 5.4. (Different colours are used for understanding purposes)

Fig 5.3



We would be using 3 different variables: Shortest path from infected bridge node (x-axis), Degrees (y-axis) and Location of node (z-axis)

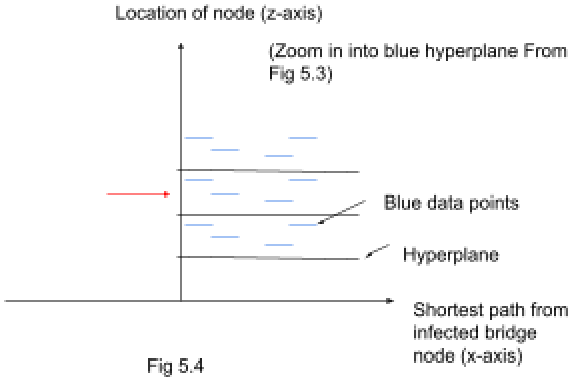

Location of node (z-axis)
(Zoom in into blue hyperplane From Fig 5.3)
Blue data points
Hyperplane
Shortest path from infected bridge node (x-axis)

Fig 5.4

Firstly, The shortest path from infected bridge node would let us see how far the node is from an infected node which would give an estimate of efficiency of giving the vaccination to that specific node as the further it is from an infected node, the less effective giving the vaccination it would be as the virus would already have travelled through multiple nodes. Therefore a lower x value would mean higher vaccine efficiency. The number of degrees would give us an estimate of the node density around the node itself showing how much the virus would spread if that specific node contracted it. This would signify the importance of giving the vaccine to that node. Therefore, a high y value would mean higher vaccine efficiency. Lastly, the location of the node. The location of the node would be found using a predetermined algorithm . The location would be plotted using the following criteria: Furthest nodes in the infected group would have the least z value, Furthest nodes in the non-infected group would have the highest z value and the nodes in a bridge group would have a z value in between in the infected group's and the non infected group's z values.

Therefore in the simple illustration of our SVM, Fig 5.4, the group in the middle with the red arrow pointing towards it, would have the highest vaccine efficiency. This group would be zoomed in on and extracted and inputted into a similar SVM to get more exact answers on to which node to vaccinate.

This would have a significant advantage over simple linear regression models as due to the increased number of variables that can be used. As well as the ease of visualising points on this graph as they are classified. Visualising 3d graphs can be done using the matplotlib library. Outliers in the data set may reduce the accuracy but not as much as it does in linear regression.

Bonus Question:

One major problem in the society that is similar to Duovid is the spread of fake news, especially through social media. Fake news is false or misleading information presented as news. It has been used to influence politics and promote advertising, but it has also become a method to stir up and intensify social conflict. Fake news online is spread through social connections between people, such as Instagram, Twitter, and other social media platforms or through news articles, advertisements and face-to-face interactions. A social connection would be defined as a medium through which data or information can be passed from one person to another on a social platform. Social connections can be represented by edges in graphs whilst people would be represented by nodes. The social network can be represented in a graph similar to the one of Duovid. We chose the problem because its process of spreading is similar to Duovid (However, while Duovid needs 2 infected people to give a person the virus, fake news only needs 1 person to spread the fake news) One may receive the fake news when he/she talks or interacts with someone else who already knows the fake news, be it through social media or face-to-face interactions. Additionally, similar to vaccinating someone against Duovid, we can warn or ban someone on the social media platform if he/she is known to spread a lot of fake news. Therefore, we can use similar methods as in Duovid to find out how the fake news is spreading and who to warn/ban to reduce the spreading of the fake news.

Through our knowledge of machine learning, we can develop an algorithm to track down people who cause the spread or are the connections between the spreader and the people with lots of connections. By integrating our algorithm into almost any social media platform, we could potentially send a warning to these users or ban them from the platform entirely. This would reduce the spreading of fake news, and eliminate those who are actively spreading the fake news on the platform. For example, using a code similar to 4b, we can find out who is the biggest cause of the spread of the fake news. This could help us determine who should be banned/warned, and thus would minimise the number of people receiving the fake news.

While a method to track the connections of people in social media exists to be able to remove those who are a threat to the spread of fake news, such methods to track face-to-face connections currently do not exist. Society might think that the government tracking face-to-face interactions is a breach of privacy, and would not be socially acceptable. Meanwhile, when you download a social media app, you have to agree to the fact that you would be constantly tracked and your information would be collected, therefore it is not a problem to keep track of the interactions between users. While we do not have a solution to track the spread of fake news through face-to-face interactions, we might be able to figure out such a solution in the near future. However, with the technology that we have at this point in time, we can only use the methods we have learned about Duovid and implement them to learn how fake news spreads through social platforms such as social media. A similar method may also be used to study the spread of computer viruses.

References:

1. “The Danger of Fake News in Inflaming or Suppressing Social Conflict.” *Center for Information Technology and Society - UC Santa Barbara*, https://www.cits.ucsb.edu/fake-news/danger-social.