Aalto-yliopisto
Aalto-universitetet
Aalto University

CS-E4840
Information Visualization

Assignment 3

Aitor Urruticoechea Puig
aitor.urruticoecheapuig@aalto.fi
Student N°101444219
April 2024

Contents

List of Figures

Acronyms

HIV Human Immunodeficiency Virus. 1, 8

MMDS Metric Multidimensional Scaling. 1, 6, 7

n-MDS non-metric Multidimensional Scaling. 1, 4, 5

PCA Principal Component Analysis. 1–4

Exercise 1 - Mystery data

Aalto University professor Sam Salabim spends all his spare time investigating paranormal phenomena and occult rites of the past. On a desktop computer in an abandoned vicarage, he comes across a csv file, Mystery.csv. *Download this file from the MyCourses page. It contains* $n = 1000$ *data points representing a point cloud in three dimensions (i.e., the three columns give the* x, y, *and* z *coordinated respectively). This exercise aims to study the shape of this data set by embedding it into one and two dimensions. Your answers to [the subsections] below should briefly explain what you have done.*

1.1 Trellis visualization

*Make a Trellis (small multiples, similar to Exercise 4 in Assignment 1) of 2-dimensional scatter plots of the point set. Also colour the points with a continuous scale indicating the row number (*1 *to* 1000*) of the data points.*

Figure 1 showcases the required Trellis visualization. Critically, it shows that the row number does play a role; and can be understood as a "time"-like dimension; as it seems to be the explanation of a movement pattern in the data. Further analysis will be, however, required.

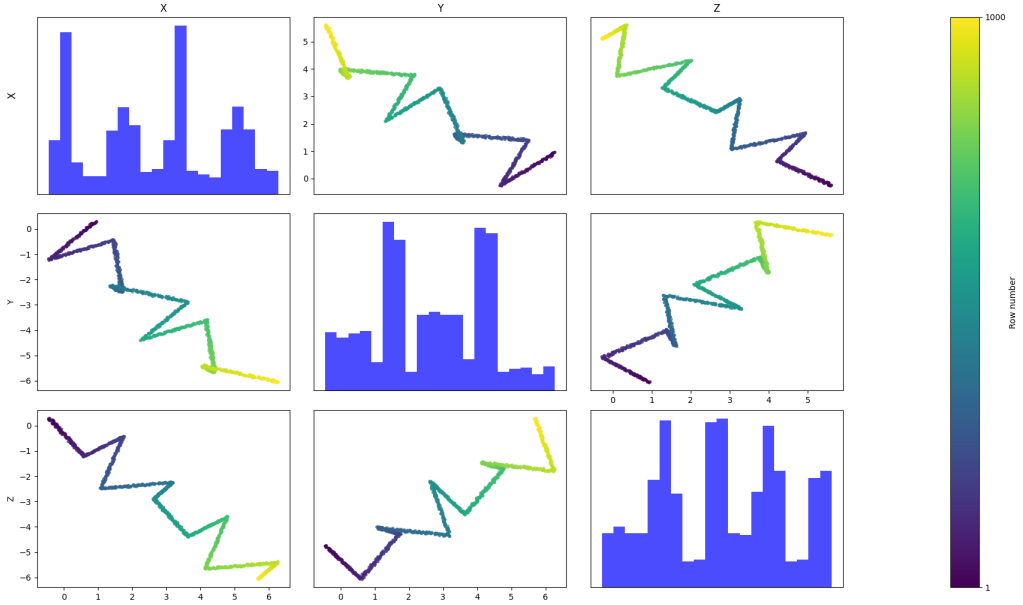


Figure 1: Trellis visualization of the mystery data.

1.2 One-dimensional PCA

Use Principal Component Analysis (PCA), project the data to the first principal component, and plot the data in one dimension using the same colour scale as in [subsection 1.1] above. Also, make a histogram of the one-dimensional embedding. With PCA, it is a good idea to centre the data first. Why? What would happen if the data [was not centred] when looking for a maximum variance projection?

Centring the data is very much a necessity in PCA. This technique is, after all, quite sensitive to the mean of the data and thus, not centring the data would easily bias the resulting PCA performed on it.

Variables with larger means may dominate the first few principal components simply because of their scale, rather than their actual contribution to the variance of the data. Then, rather than reflecting the actual structure of the data, the PCA merely reflects the mean biased by these variables. This can dangerously lead to misinterpretations of the data and even inconsistent results if different analyses are done in parallel with the same data.

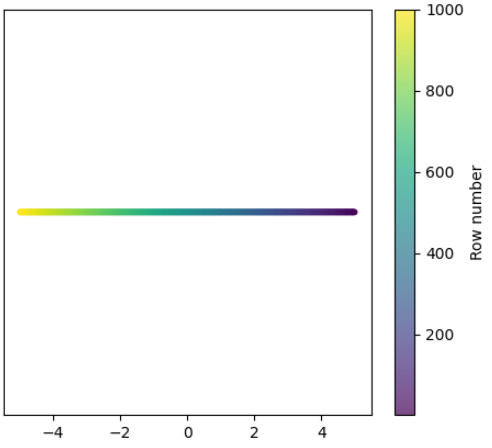


Figure 2: One-dimesnional PCA of the data.

Acknowledging this need, the data has been centred, and the data has been projected to its first main component, as can be seen in Figure 2. As previously predicted, the time-like component represented by the row number seems to indeed be the principal component that defines the behaviour of this mystery data. Interestingly, the plotted histogram (Figure 3) showcases quite a consistency in the distribution of the data throughout this principal component, with very similar frequencies for each bin, to the point where the differences between them have been exaggerated in the visualization to allow for the observation of a few very small valleys. In these, the first principal component frequency drops ever so slightly to the 47 and 48 frequency marks.

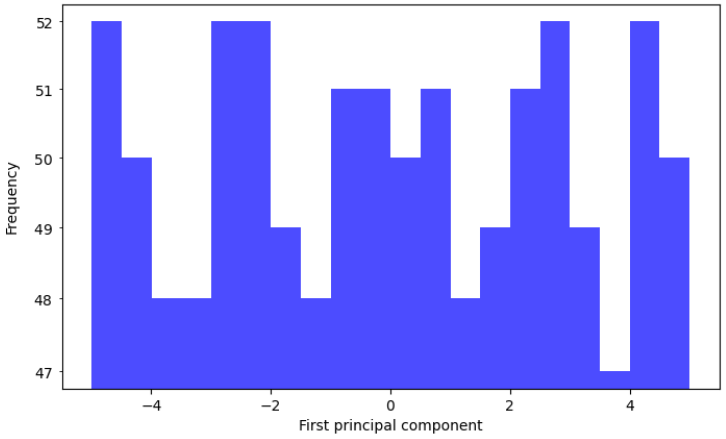


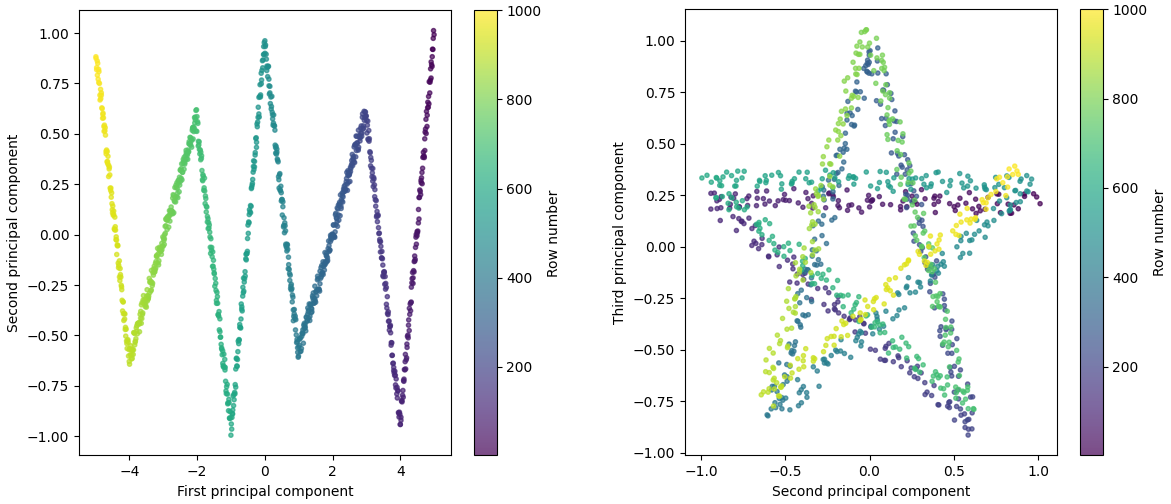
Figure 3: Histogram of the first component of the mystery data.

1.3 Two-dimensional PCA

Then make two-dimensional plots of the data projected to the (plane defined by the) first and second PCA components, and to the second and third components. Based on these, what can you tell about the data set's shape?

Figure 4 showcases the two required scatter plots. The one resulting from the combination of the first and the second principal components (Figure 4a) gives little extra information apart from being a cleaned-up version of the Trellis plots previously performed (Figure 1). Though the peaks are cleaner and the lines are displayed in a more straight form, it is hard to extract extra information other than the fact that the suspected time-like dimension of the row numbers is still behaving as

such. A very interesting phenomenon happens, however, when looking at the plot of the second and third principal components (Figure 4b). There, the data aligns itself into a pentagram-like pattern; forming a star shape. Even more fascinatingly, when looking at the time-like dimension represented by the row numbers, it is of relevance to see how, by row number, the different data points seem to gradually draw the pentagram. They end up drawing it exactly twice, starting (with the point in row 0) and finishing (with the point in row 1000) in the same vertex at $(1.0, 0.25)$. This is too much of a concrete pattern for it to arise randomly; and the author of the data clearly wanted this to appear when or if anyone were to ever try to decipher the meanings of the file.



(a) First and Second principal components.

(b) Second and Third principal components.

Figure 4: Two-dimensional PCA analysis of the mystery data.

1.4 n-MDS plot

Use non-metric Multidimensional Scaling (n-MDS) or Sammon mapping to embed the data into one or two dimensions and plot the data as you did in [subsection 1.2] above.

A n-MDS plot with two dimensions has been chosen for this exercise (see Figure 5). However, this does not seem to help much in deciphering the behaviour of the data. The scatter plot is now as scatter as it gets; and little new conclusions can be extracted. However, some new, albeit of debatable usefulness, information is indeed now visible when looking at the time-like dimension of the row numbers. It seems like the dispersion of the data changes with this row number dimension. Starting somewhat centred with the first rows very much in the centre of the scatter plot; and as the row number increases (or time passes if one chooses to interpret it this way), it becomes more scattered, revealing the need for a more zoomed-out scatter plot.

1.5 ISOMAP plot

Use ISOMAP *to embed the data into one or two dimensions and plot the data the same way you did in [subsection 1.2] above. Hint: A good definition of the neighbourhood is to require that items* i *and* j *are neighbours if* j *is one of the* k $(k < 10)$ *closest points to* i *or if* i *is one of the* k *closest points of* j *(but you can also use some other definition of neighbourhood).*

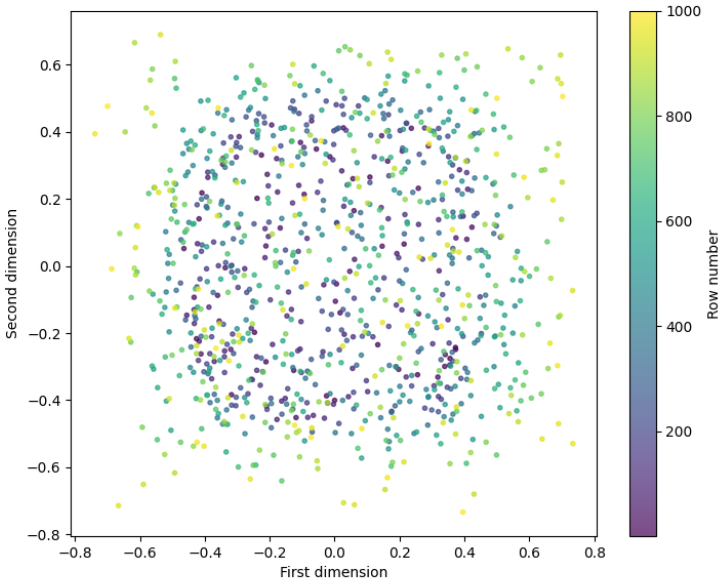


Figure 5: n-MDS two-dimensional plot of the mystery data.

Following the recommendation; a k value of 7 has been chosen to limit what is considered a neighbouring data point. With that, Figure 6 has been obtained, which seems to imply a much simpler data distribution than what has been shown so far. A fairly unremarkable normal-like distribution spanning from -10 to 10 in the first dimension (after data centring) with a few outliers that ultimately do not compromise the actual pattern of the figure.

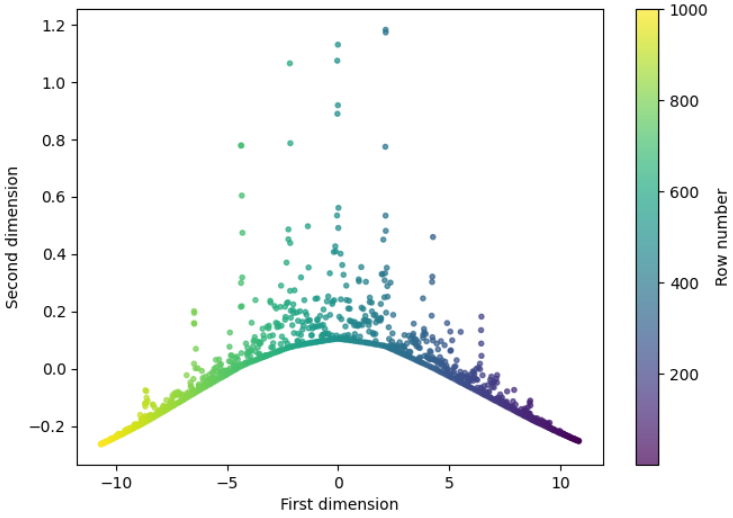
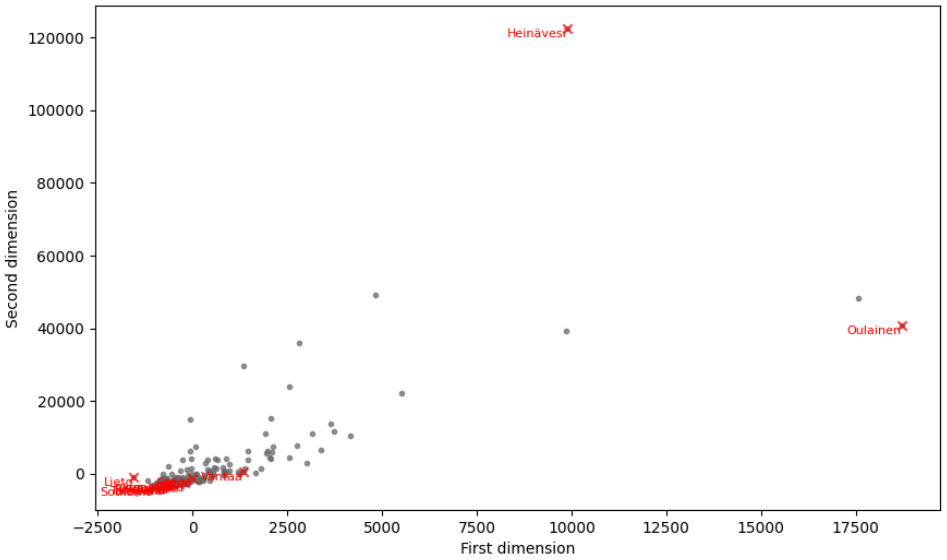


Figure 6: Isomap two-dimension plot of the mystery data.

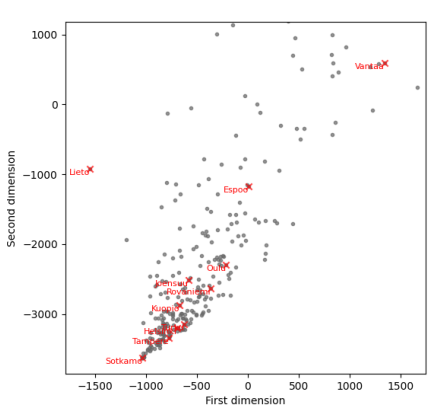
Exercise 2 - Finnish population age distribution

Download from MyCourses the dataset population_data.csv *which contains statistics of population age structure in Finnish municipalities. The data is organized in diferent age groups with the following columns. Compute and plot Metric Multidimensional Scaling (MMDS) and Sammon mapping. Annotate some selected (or all) places, for example, main cities, provinces, places where you have been/born, etc. Compute Shepard plot (a scatter plot of output distances as a function of input distances) and compare the plots of MMDS and Sammon mapping. Which method predicts which distances better?*

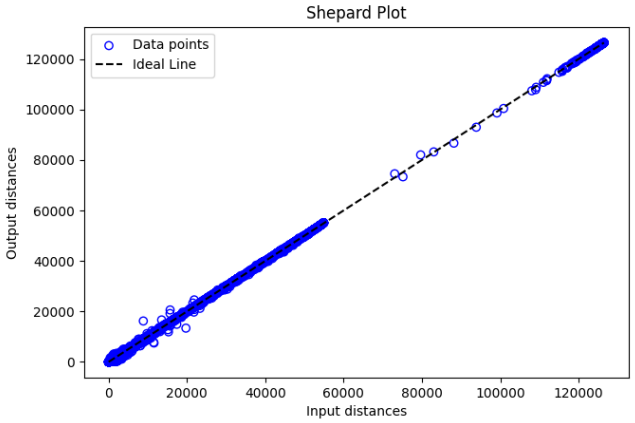
Figures 7 and 8 include, respectively, the generated MMDS and Sammon mapping of the population data; as well as the accompanying Shepard plots. In both cases, a list of relevant Finnish cities have been highlighted. These include Helsinki, Espoo, Vantaa, Tampere, Rovaniemi, Turku, Oulu, Joensuu, and Kuopio. Added to these, relevant outliers have also been highlighted. These are, in both cases, quite small towns that seem to break expected norms in population distribution precisely because of the small population they represent. It is quite interesting, however, to see how Heinävesi appears every time as the single data point that forces the zooming out of the graphic.



(a) General MMDS visualization of the population data.

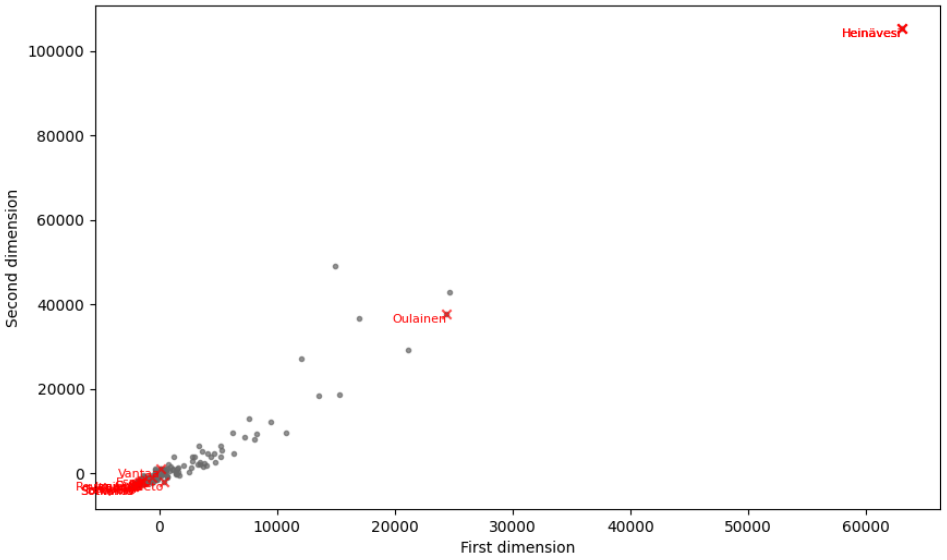


(b) Zoomed-in plot of the MMDS map.

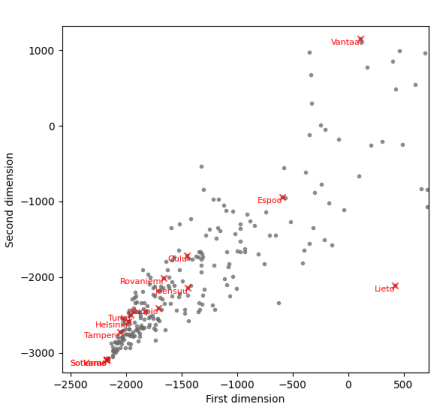


(c) Shepard plot for the MMDS map.

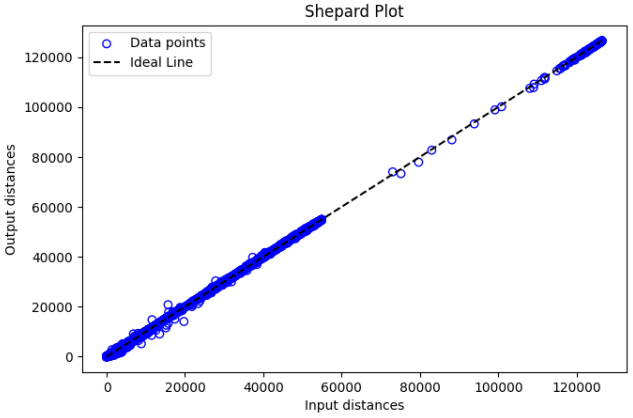
Figure 7: MMDS mapping for the population data.



(a) General Sammnon visualization of the population data.



(b) Zoomed-in plot of the Sammon map.



(c) Shepard plot for the Sammon map.

Figure 8: Sammon mapping for the population data.

Comparing both Shepard plots (Figures 7c and 8c) proves quite tricky, as both methods seem to struggle with the same data points. However, upon closer inspection, it is clear that the Sammon map manages to handle slightly better the points corresponding to larger distances. Shorter distance points are harder to analyse, but by (again) a small margin MMDS seem to do the job ever so slightly better.

Exercise 3 - Graph visualization

From Mycourses, download the dataset graphs.zip*, which contains several networks specified as adjacency lists—lists of which node that is connected to which other node with id-numbers ranging from* 0 *to* $N - 1$*, where* N *is the number of nodes. In the zip file, there is a readme file with further descriptions. Pick one of these networks. Visualize it using the principles introduced in the last lecture (and the general Tufte's principles taught earlier). Explain why your visualization is appropriate for this network and how you produced it. Also, visually indicate each node's given attribute labels and try to make different network substructures visible. You may use any software (e.g., yEd, PowerPoint, Illustrator, etc.) or even hand drawing to develop and present your solution.*

The Human Immunodeficiency Virus (HIV) hit hard in the 1980-decade United States of America; and affected specially the queer population. Uninformed and marginalised, information about the importance of blood and sexual transmission nature of the disease was only widely distributed years too late. Figure 9 presents one potential spreading pattern for the first group of infection in North America. Using a Kamada Kawai force modelling for distributing nodes and connections, this seemingly clean graph has been generated. This is quite the appropriate approach because Kamada Kaway allows for minima-to-no connection crossings, and more or less equal spacing between nodes; which makes a lot of sense when trying to visualize the spread of a disease. The graph has then been coloured according to the number of connections each node has. Blue represents ends of the transmission chain, where affected individuals did not spread the disease further. Purple represents connection individuals, who by (bad) luck of the draw managed to bring HIV from one individual to the next. Orange is used for "adder" individuals, who having been infected by one of their connections ended up spreading HIV to two others. Finally, red is used for "spreaders", for those individuals that ended up in a position where their role managed to be a catalyst for major multiplication of HIV-affected individuals. It is important to note that this graph and subsequent analysis does not aim to place blame on any of the victims of HIV, who with almost full certainty performed their role in this graph with full ignorance of what was truly happening. Far from this, graphs like this one can help researchers in understanding pandemic patterns, propensities that individuals may have to contract or be immune to the disease, etc.

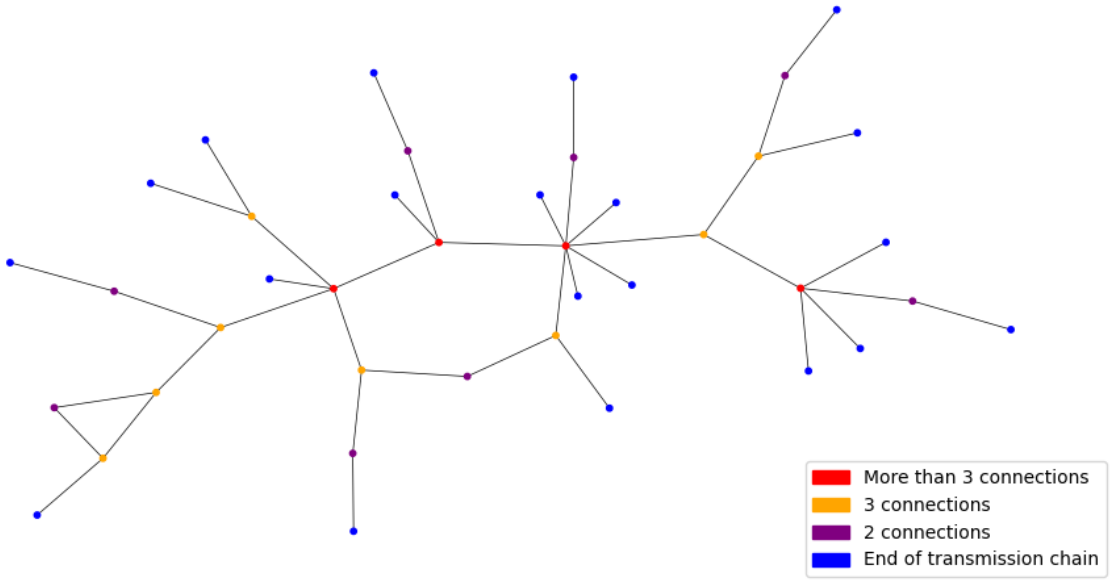


Figure 9: Potential first infection pattern for HIV.