Mapping HIV Transmission Networks for Targeted Prevention Policies

Daniela Baraccani, Artificial Intelligence, 0001038306
Afshin Khodaveisi, Artificial Intelligence, 0001050589
Guido Laudenzi, Artificial Intelligence, 0001033343

1 Introduction

In the field of virology, the study of the Human Immunodeficiency Virus (HIV) remains an open research due to its significant implications for public health and its complex interactions within human populations. The HIV virus, which compromises the immune system and can lead to Acquired Immunodeficiency Syndrome (AIDS), presents a dynamic challenge that intersects with diverse domains such as genetics, epidemiology, and social sciences.
This project is posed at the intersection of virology and network science, aiming to utilize the analytical capabilities of NetworkX, a Python library designed for the investigation of complex networks. By exploring a dataset comprised of a set of different HIV test results, this study seeks to understand the patterns and metrics inherent in the spread of HIV. The specific application of this project involves a quantitative analysis of the network characteristics of HIV transmission to better understand the structural properties of the virus's propagation within a population and to possibly give some idea to act upon them.

2 Problem and Motivation

Despite the availability of effective treatments, HIV continues to affect approximately 39 million individuals worldwide as of the end of 2022, with annual fatalities exceeding 600,000 due to late diagnosis, inadequate treatment, or lack of access to healthcare services [1].
The central aim of our project is to unravel the intricate network structure of HIV to inform and enhance prevention strategies. By conducting a comprehensive network analysis, we intend to deepen our understanding of HIV's spread, identify high-risk nodes within social and sexual networks, and uncover patterns of contagion that have previously gone unrecognized. This aspiration is not solely for academic enrichment but carries significant implications for shaping public health policies and providing individual risk assessments.
Through this analysis, we want to pinpoint the principal vectors of HIV transmission and assess the most effective social policies that could meaningfully reduce the virus's transmission rate. The ultimate goal is to transform raw data into actionable intelligence, contributing to the global fight against HIV by preventing new infections and reducing the mortality rate associated with this virus.

[1] World Health Organization on HIV

Other works on HIV using network analysis exist. One of the most influencial papers [1] aims to comprehend the patterns of transmission by mapping out the relationships and interactions between infected individuals among rural drug users. It explores how these networks influence the spread of HIV, shedding light on the dynamics that contribute to its transmission within a certain population; another result [2] studied a group of 96 HIV-positive patients through network stability, and changes in risk configuration and risk behavior using epidemiologic and social network analysis.

3 Datasets

The datasets for our project were obtained from the Colorado Index of Complex Networks (ICON) [2], an exhaustive repository of network datasets utilized across various domains of network science. These domains range from social and biological to transportation and technological networks. Our chosen datasets [3] specifically track HIV transmissions, delineating the paths of the virus through sexual contacts, needle-sharing practices, or broader social interactions. This comprehensive dataset unifies eight distinct studies conducted between 1988 and 2001, offering a rich historical perspective on the HIV epidemic's evolution over time.
To ensure a robust and detailed analysis, individual datasets contain not only HIV test results but also a spectrum of other diseases, supplemented by demographic details like age, ethnicity, and gender. The networks are constructed as directed graphs with unweighted connections, capturing the directionality of potential transmission without weighting for frequency or intensity of contact. Overall, the network comprises 7132 nodes and 27588 edges.
Our dataset is divided into two types of networks: egodyads, which represent connections between study participants and their immediate partners, and altdyads, encompassing individuals linked to the primary egodyad network. However, for the purposes of our project, we will analyze them as a unified entity to maintain the integrity of the transmission dynamics.
In terms of data handling, the dataset is publicly available through ICON, ensuring transparency and reproducibility of our research. For computational measures, NetworkX serves as our primary tool, enabling us to calculate network properties such as small-worldness, degrees of connectivity, centrality, and clustering coefficients. The use of NetworkX is instrumental in translating the complex network data into insightful metrics that can inform public health strategies and interventions.
The HIV dataset includes 36 features covering various aspects, initially including unrelated illnesses like chlamydia, which have been removed to focus specifically on HIV-related factors. Within this dataset, the presence of missing values is prevalent across multiple features, with some instances reaching as high as 80%.

4 Validity and Reliability

The validity of our study is based upon the fidelity with which our model represents the real-world network of HIV transmission. By selecting nodes that are HIV-positive and their directly connected counterparts, our model aims to reflect the immediate social and relational context of the virus's propagation. This approach provides insights into not only the transmission dynamics but also the social structures that may facilitate the spread of HIV.
In terms of reliability, the treatment of data and the criteria for node inclusion are systematic

[2] ICON Website

and transparent, ensuring that the study can be replicated and the findings reproduced.
In particular, the shrinkage of the network maintains a balance between a comprehensive representation of the transmission potential and the manageability of the dataset, creating a smaller subgraph of only 175 positive nodes (graph structure in Appendix, Figures 1, 2). The reproducibility of the study is confirmed by these algorithmic approaches to data treatment, allowing other researchers to apply the same methodology to similar datasets and compare results.

To enhance the robustness of our study on HIV transmission networks, we could have considered alternative modeling approaches, each with its own set of implications.
Firstly, incorporating all nodes, regardless of HIV status, would have yielded an extensive network, encompassing both HIV-positive and negative individuals and their myriad connections. This inclusive strategy might have offered a more holistic view of the network's structure. However, it would likely have obscured critical insights into HIV transmission dynamics. Specifically, nodes representing HIV-negative individuals with exclusive connections to other negatives could have been misleadingly emphasized in centrality measures, suggesting a prominence in the transmission network that does not align with their actual risk or contribution to the spread of HIV.
Secondly, another interesting approach would have been to refine the network further by retaining only the direct connections between HIV-positive nodes and their HIV-negative counterparts, effectively pruning multiple-degree separations. This method could potentially produce a less fragmented network, potentially revealing interconnected clusters or cliques that signify direct risk pathways. Nonetheless, while this might simplify the network and highlight immediate transmission risks, it would also diminish our ability to trace and understand the patterns of HIV spread among positive individuals. By not capturing the indirect connections that might eventually lead to new cases, this model would provide a limited perspective, underestimating the complexity of the transmission network.
Our chosen methodology, by focusing on HIV-positive nodes and their direct connections, seeks a middle ground that balances detail with clarity. It aims to construct a manageable yet informative subgraph that prioritizes the transmission potential to its roots.
At the end, the second modeling approach is used to inspect how policies estimated with the positive-only graph can effectively prevent infections among negative patients.

5 Measures and Results

5.1 Data inspection

In order to find a subset of nodes' features for the network analysis, we employed a robust methodology utilizing `sklearn`'s Random Forest [4] algorithm to discern crucial patterns among interconnected nodes. With an extensive dataset of 36 features, our approach centered on identifying the most influential variables for analysis. Through a strategic voting mechanism integrating various random forest formats, we selected a subset of 5 key features: patients' race, engagement in sex work, involvement in drug dealing, sexual orientation, and disability status. These features were judged as fundamental in understanding the network dynamics, as they significantly contribute to predicting tie types between nodes, as well as discerning positivity within nodes and their connections. This approach not only showed the complexity of the dataset but also enabled us to focus on the most pertinent attributes driving the network's structure and composition. Features' importances are shown in the Appendix (Figures 3, 4, 5, 6).

5.2 Centrality Analysis

Centrality measures in network analysis are crucial because they help identify the most influential or important nodes within a network. They quantify the relative importance of nodes based on different criteria, shedding light on various aspects of network structure and functionality.
We performed various centrality measurements: firstly we performed Degree Centrality, followed by Eigenvector Centrality, Katz Centrality, Closeness and Betweenness Centralities. There were predictably some differences on their node rankings, but also a lot of similarities, i.e., node with ID 29 was overall the most important one for all the measures.
As mentioned in the previous chapter, since the nature of the dataset, a lot of questions were left blank, such as sexual orientation or drug dealing. On the other hand, we found that the edge's tie type feature of our dataset did not include any missing value. In addition to this, given its crucial position for deducing possible policies for HIV prevention, we decided to consider it as the most important focus during our study of HIV transmission.
We inspected the distribution of the top 5 or 10 nodes' features for each centrality measure, and though they slightly differ from centrality measure to measure, they highlighted some common interesting facts: the top 5 nodes in every centrality ranking show that every relationship between nodes is characterized by needle sharing. Analysing the top 10 most central nodes, there is just one case of a sexual relationship. This means that HIV transmission for the most important nodes has likely occured through needle sharing, possibly caused by a drug addiction background.
Other factors that highlight commonalities between all the most central nodes are: the age of subject, their race and their gender. We found that all top nodes are male subjects, mostly in their 30s, ranging from 29 to 42, with an outlayer of 92 (which we attribute as a collection error, since assigning 1900 as year of birth seems to be a placeholder).

5.3 Group of nodes

Cliques The network comprises 126 cliques, with only 18 composed of three or more nodes. In particular, two 4-cliques are present, of which we studied the features, with the aim of observing analogies. The investigation of cliques and their characteristics reveals that, overall, cliques with three nodes or more exhibit stronger correlations (features sharing) between nodes. The primary suspected reasons for virus transmission within these groups are social interactions and drug sharing, opposed to the sex connection which never happears at tie type. Moreover, the majority of included nodes are male (in one 4-clique, proportion of 50%; in the other 100% males), and the age range is between 28 and 45 years.
On the other hand, the examination of nodes within cliques does not indicate any abnormalities concerning the number of sex-workers and drug dealers, and their impact on HIV transmission between related nodes. It is important to note, however, that a considerable number of individuals have not provided information regarding these features, emphasizing the need for careful consideration of this significant lack of data information.
Regarding participant orientation, one clique comprehends a bisexual subject; the all-males clique includes three homosexual subjects even if their tie type is drug sharing.
Additionally, a noteworthy remark is that a significant portion of affected individuals are of white and black ethnicity (in one 4-clique, all blacks), even if the overall percentage of Asian individuals in the dataset is higher than that of white/caucasian individuals.
Finally, concerning individuals with disabilities, it's intriguing to mention that half of those

who shared this information reported having a disability.

Cores In analyzing the cores (or groups of nodes) by applying the k-core measure and sampling multiple values for k, it was observed that the group with the highest k (degree of nodes) is 3, as shown in Figure 13. There are two cores under these circumstances which are represented as full density graphs (the previously mentioned cliques).
Furthermore, the largest group with a degree of two is the cluster that includes node 29, which is the most central node of the network according to all the centrality measures previously inspected (Figure 12).
Most of the features values are similar to the ones of the biggest cliques of the network.

5.4 Community Detection and Clustering

Triangles Upon analyzing triangles - subgroups of three connected nodes in the network - we unsurprisingly found that the nodes heavily involved in these configurations (nodes that are a part of the highest number of triangles) are also the most central ones. Once more, node 29 emerged as the most prevalent within triangles, reaffirming the outcomes derived from the centrality measures. Notably, the dominant factor remains the prevalence of needle sharing relationships, suggesting that HIV spread within the triangle relationships in our network primarily through needle sharing.

Transitivity Transitivity can be defined using the sentence *friends of friends are friends*. Our network exhibits a very low transitivity coefficient (0.19), indicating that it is not tightly interconnected as a community. In the context of our study, this is advantageous because tightly knit groups often facilitate faster virus transmission.

Clustering coefficients The clustering coefficient of a single node measures the proportion of connections between its neighboring nodes out of the total possible connections among them. It reflects the *cliquishness* or tendency for nodes to form clusters or tightly knit groups. While it is worth pointing out that subjects with a clustering coefficient equal to 1 are also the nodes appearing in cliques or cores, the overall average clustering coefficient of the graph is only 0.16: it indicates that, on average, the immediate neighbors of nodes in the network form connections with around 16% of their potential connections. In other words, subjects exhibit some tendency to form clusters or groups but not extremely densely interconnected ones.

5.5 Path Analysis

Given that our network is disconnected, we have identified 44 sub-graphs linking the nodes together. The largest sub-graph comprises 62 nodes, while the remaining ones exhibit lower degrees. Specifically, in descending order, there are two paths with six nodes each, followed by two paths with five nodes each. Therefore, our network contains a significant connected component alongside numerous smaller subgroups. The disconnection within the graph can be seen as an advantageous feature, simplifying the isolation and analysis of major groups, especially when studying virus transmission.

Within the 62-node subgraph, the edges are predominantly characterized by the needle sharing tie type. Specifically, we observed 76 edges associated with it, 3 edges linked to sexual ties, and 4 edges designated as other ties. The ages of the individuals range from 25 to 44, with 30 and

40 being the most frequent values (excluding outlier values such as 92). Regarding ethnicity, 49% of the subjects identify as black, 15.8% as white, while the remaining subjects did not provide an answer. Regarding gender, 44% are male, 16% are female. All of the subjects in this group belong to the study made in Brooklyn, the *Bushwick SFHR*.
In the first 6-node subgraph is characterized by needle sharing related ties. 50% of patients are black and the other half have not provided an answer; all of the subjects are male. The members of this group belong to the same *Bushwick SFHR* study.
In the second 6-node path all nodes are linked through drug tie type and some of them also through sexual relationships, in addition to the 'drug' tie. Three of them have a disability. The subjects of this group are a part of the study carried out in *Houston*.
The first 5-node subgraph has again only drug-related connections. 4 of the 5 nodes in this path are black men, and one has not provided an answer regarding their race nor their gender. Lastly, two of these subjects have a disability. The subjects of this path belong to the *Atlanta Urban* study.
In the second 5-node subgraph have needle-sharing relationships, and an inner triangle of nodes have also a sexual relationship. In this path, there are only young subjects, such as 24 or 27 years old, the oldest being 36. All of the subjects in this path are white men, three of them have a disability, and they are all subjects of the study made in *Houston*.

For concluding: all relationships have a common presence of drug/needle-sharing ties; the affiliation of the subjects to a different path and study implies a probable common geographical location, or context (which could influence risky behaviours); the demographic variety highlights a complex subjects profile. Moreover, the second 5-node path shows a more complex transmission dynamic because of the presence of sexual ties, bringing light to another factor for possible intervention in case of this specific group; the varying age of the subjects can also give us information on the risk factors, highlighting the importance of an early intervention for at risk young drug users.

5.6 Homophily and Assortative Mixing

Our network demonstrates a near-zero degree assortativity value (in particular, 0.1332), equal to the Pearson correlation coefficient. These nearly null assortativity coefficients signify that subjects do not significantly link with others who share similar or dissimilar characteristics. It portrays a somewhat *heterogeneous* network in its relationships. This assortativity value suggests that there is not a clear pattern of homophily (preference for similar connections) or heterophily (preference for dissimilar connections) in the network. This can have implications for identifying distinct subgroups or communities within the network, as there might not be well-defined clusters based on shared attributes. It implies that prevention efforts might be more effective if focused on nodes with high centrality rather than on groups of nodes with similar profiles.

5.7 Small-worldness

The concept of small-worldness in network analysis refers to the balance between high clustering (like a tightly knit community) and short average path lengths (the degree of separation between any two nodes). A network is considered to have the small-world property when it exhibits a high level of local clustering (nodes tend to group together) and yet the average distance between nodes remains relatively small. Our findings of *sigma* and *omega* values of 1.6

and -0.15, respectively, indicate that the largest subgraph of 62 HIV-positive nodes possesses the small-world property: this suggests that while there might be distinct nests within the network (possibly representing different communities or groups), there are also relatively short paths connecting different nodes. This interconnectedness could imply efficient information or disease transmission routes between nodes, highlighting potential avenues for intervention or targeted outreach within these clusters to impact the broader network of HIV-positive and negative individuals.

5.8 Inferring possible policies to prevent probable transmissions

Upon observing the network comprising solely HIV-positive subjects, we have identified needle sharing among individuals that are possibly struggling with drug consumption/addiction as the most prevalent method of transmission. Consequently, we hypothesize that an effective and impactful approach to reducing disease transmission would involve providing clean needles or giving greater consideration (potentially through increased access to free HIV tests) to individuals within this demographic.
Our analysis reveals a pattern similar to profiling within the network. Various factors contribute to shaping a shared background among HIV-positive individuals of this graph. These factors encompass characteristics such as disability, sexual orientation, race, gender, and age. It suggests that certain demographic and personal attributes might play significant roles in the HIV-positive population. Understanding these shared characteristics can offer valuable insights into the correlation of these factor and could help in designing more targeted interventions to address the needs of these specific subgroups within the population.

In our concluding study, we conducted a brief exploration of the connections between our network and the one comprising the first-degree HIV-negative subjects directly linked to HIV-positive individuals, aiming to understand the potential implications of our proposed social policy. Among approximately 230 nodes in this initial negative layer, about 200 of them have direct connections to HIV-positive individuals due to drug-related ties. Then, we focused on examining the connections of the 10 among the most central nodes and the two 4-cliques within this subset. Remarkably, these nodes and cliques account for a significant portion, precisely 184 connections, highlighting their crucial role in the transmission of HIV. This finding underscores the critical importance of central nodes in the spread of HIV within the network. Identifying and targeting these central nodes, which act as key influencers or bridges between different subgroups, becomes imperative when devising intervention strategies. By concentrating efforts on these influential nodes, we can potentially disrupt the transmission pathways and mitigate the spread of the virus more effectively.

6 Conclusion

Our research focused on unraveling the patterns of an HIV transmission network, aiming to identify effective intervention strategies through social policies to decrease, or ideally halt, the spread of the HIV virus.
Our study notably underscored the drug user community as the most vulnerable group to contagion, primarily through the use of infected shared needles. We arrived at this conclusion by employing various measurements on individual nodes and node groups. This involved scrutinizing the characteristics, profiles, and relationship types of the most central nodes, alongside

exploring measures such as cores, assortativity, and small-world properties.
The specifics of the spread can be summarized as a rapid event, as indicated by the small-world property, where central nodes play a more substantial role in disease dissemination than the influence of specific profiles or characteristics of clusters. Further evidence lies in the fact that key nodes, particularly those forming the largest cliques, are predominantly the most connected to the first layer of negative nodes. This insight highlights the critical importance of targeting these influential nodes when applying the proposed policies. Concentrating efforts on these key connectors within the network could significantly disrupt transmission pathways and consequently mitigate the spread of the virus more effectively.

7 Critique

While discoveries are surprising, there are limitations within the dataset that impact the quality of our results. Improved transparency from participants, particularly regarding details about drug addiction, occupation, and sexual orientation, would have significantly enhanced the accuracy of our findings.
One major concern lies in the sincerity of the provided information. For instance, there are specific cases in the dataset where some answers given by subjects about the nature of their relationships with other subjects are not reciprocated or even contradictory.
Obviously, discussing sensitive topics like HIV transmission can pose cultural barriers, leading to approximately 80% of subjects either refusing to answer questions or potentially providing less than truthful responses. This reality significantly affects the reliability of our dataset and necessitates cautious interpretation of our results. Furthermore, the dataset exhibits a bias, primarily with most connections being attributed to drug-related ties. This bias limits our ability to draw direct comparisons between drug-related and sexual transmission.
Another thing to add is that the nature of the questions asked during data gathering is often not valuable, and some data that would have been beneficial to add to the dataset has not been asked. For example, it would have been a lot more useful, for the aim of our study, to know if a subject is a drug addict, rather than a housewife or a drug manufacturer, since transmission is more related to the use of an infected needle.
These limitations underscore the need for a kind interpretation of our findings. It's crucial to acknowledge the potential inaccuracies coming from the dataset's limitations and the cultural and social complexities surrounding HIV-related matters. Despite these challenges, our study provides valuable insights, albeit within the context of these constraints.
While they serve as a starting point, our proposed policies are not yet confirmed to be universally workable solutions. Rather, they offer a foundational basis from which to develop more sophisticated, multifaceted policies that consider the multifarious aspects of HIV transmission within diverse communities. Efforts to mitigate HIV spread must evolve and incorporate a more comprehensive understanding of the complex network dynamics and socio-cultural influences at play.

References

[1] April Young et al. “Network Structure and the Risk for HIV Transmission Among Rural Drug Users”. In: *AIDS and behavior* 17 (Nov. 2012).

[2] RB Rothenberg et al. “Social network dynamics and HIV transmission”. In: *AIDS* 12(12) (Aug. 1998).

[3] Martina Morris and Richard Rothenberg. “HIV Transmission Network Metastudy Project: An Archive of Data From Eight Network Studies”. In: *Inter-university Consortium for Political and Social Research* (1988–2001).

[4] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

A Appendix

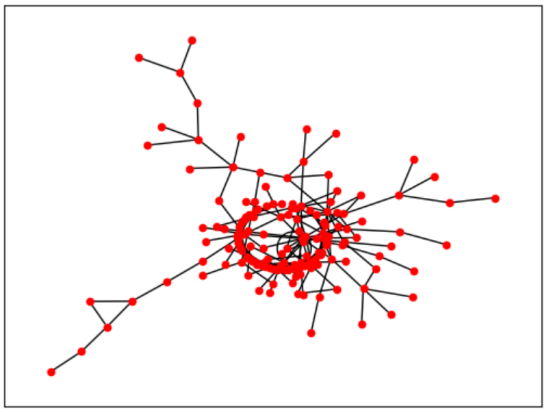

Figure 1: NetworkX HIV Graph: result is not satisfying, another Python package must be used.

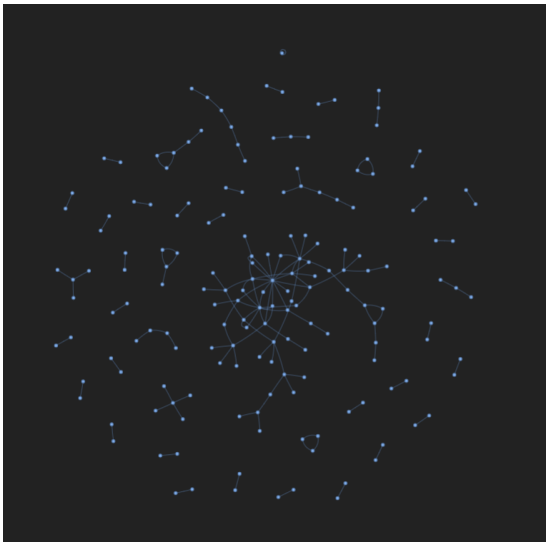

Figure 2: Pyvis HIV Network

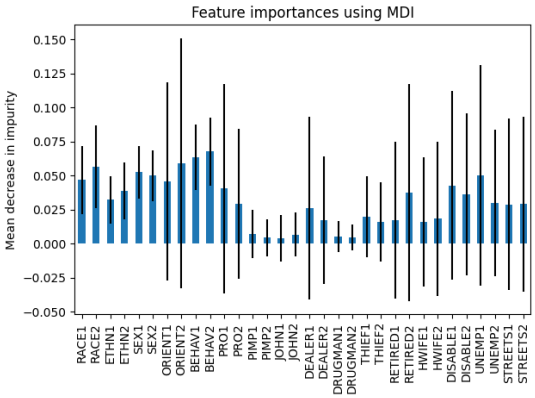

Figure 3: Feature importances using MDI: predicting tie type using both nodes features.

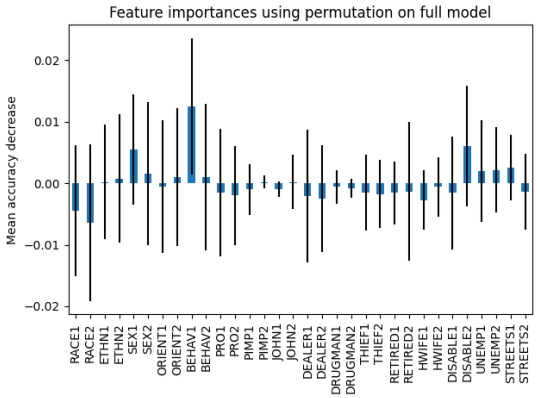

Figure 4: Feature importances using permutation: predicting tie type using both nodes features.

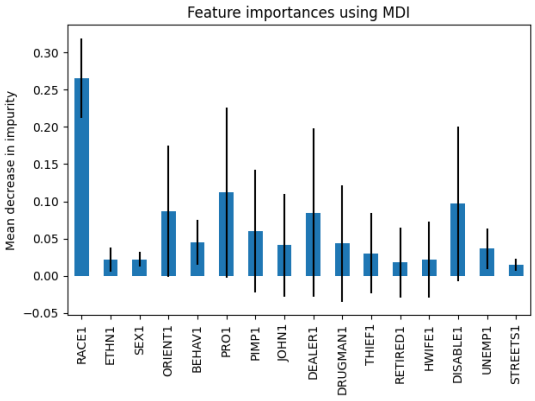

Figure 5: Feature importances using MDI: predicting HIV positivity using personal information.

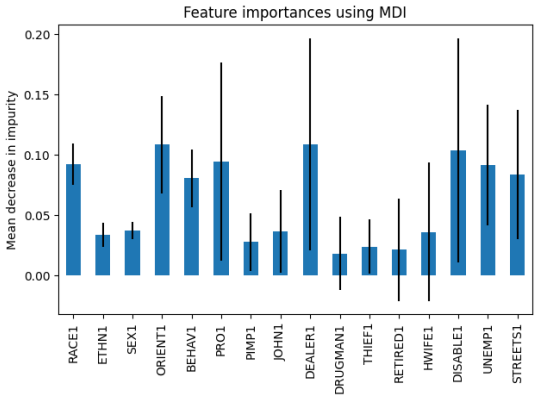

Figure 6: Feature importances using MDI: predicting HIV positivity of connected nodes using personal information.

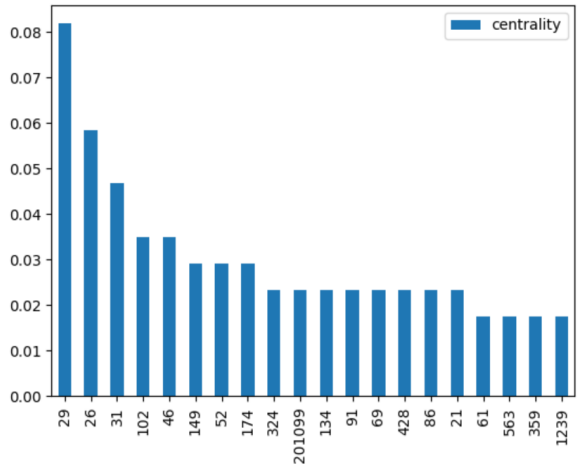

Figure 7: Degree Centrality

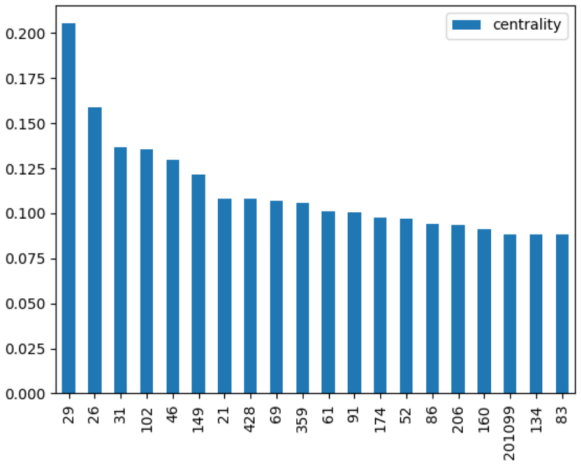

Figure 8: Katz or PageRank Centrality

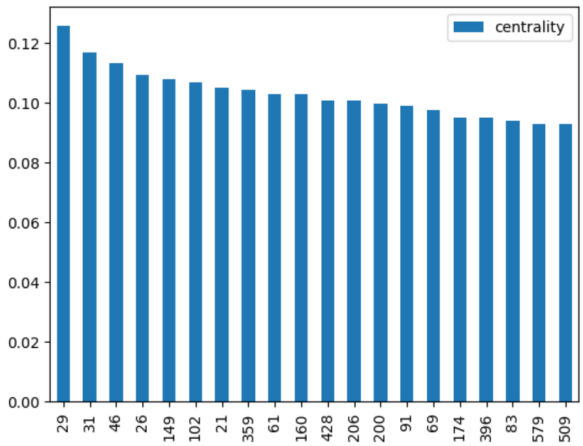

Figure 9: Closeness Centrality

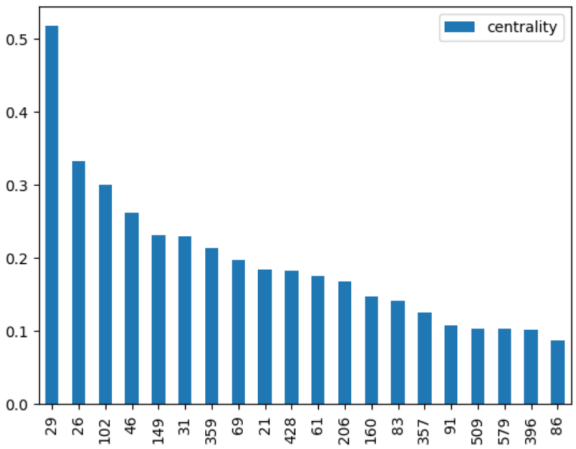

Figure 10: Eigenvector Centrality

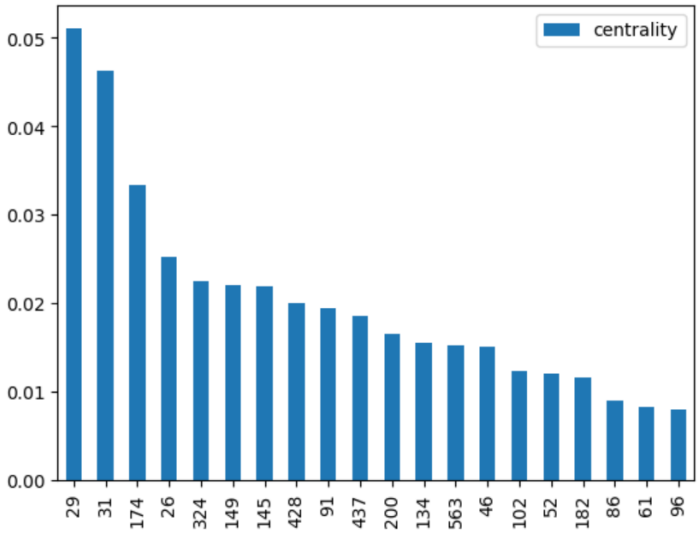

Figure 11: Betweenness Centrality

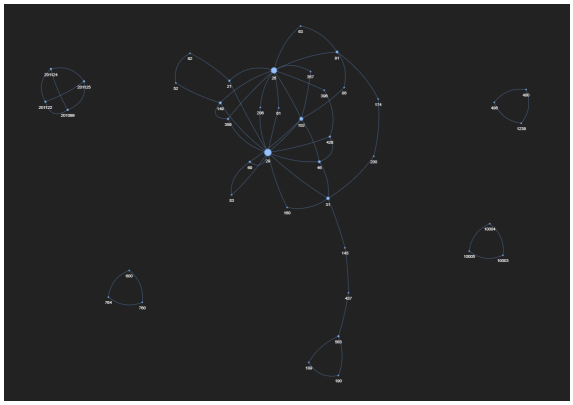

Figure 12: 2-cores

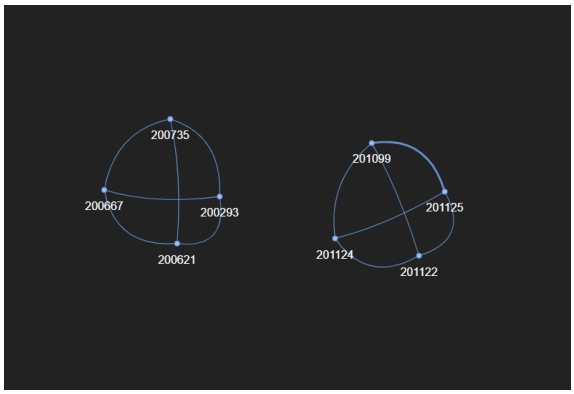

Figure 13: 3-cores or 4-cliques