## Complex networks and agent-based models of HIV epidemic

Zarrabi, N.

**Publication date**
2013

# 4

# Combining Social and Genetic Networks to Study HIV transmission in Mixed Risk Groups

This chapter is based on N. Zarrabi, M. Prosperi, R. Belleman, A. De Luca, P.M.A. Sloot, “Combining Social and Genetic Networks to Study HIV Transmission in Mixing Risk Groups", European Physical Journal, EPJ Special Topics, In press (2013).

## 4.1 Introduction

Complex Networks have provided insight in understanding many natural phenomena [162, 177], such as infectious diseases spreading [56, 152, 107], financial markets [23, 132], transportation and mobility in new socio-technological systems [173, 4] and so on. The complexity of such systems arises from dynamics at different spatio-temporal scales, from molecular to individual level dynamics, all the way to the population level [185, 150]. For example, infectious disease spreading in a population is dictated by social interactions as well as the genetic diversity of the transmitted viral agent at the molecular level. Understanding the multi-scale nature of such complex system requires data from both micro-level (i.e. genome sequencing or individual behaviours) and macro-level dynamics (i.e. social interactions or the distributions and community structures in human societies). Therefore, combining data from different spatio-temporal scales is crucial in order to better understand the behaviour of a complex system. Human immunodeficiency virus (HIV) transmission as a particular application of infectious disease spreading is studied for more than 30 years [125, 137, 104]. Sociological, clinical and virological data have been gathered from HIV infected patients. These data are from different scales such as sexual behaviours and promiscuity of infected individuals and viral genomes sequencing and mutations rates of viral strains [125, 137]. At the molecular scale, the genome sequencing data indicates the genetic sequences of viral strains that patient are infected with. Molecular virologists use these genetic information and apply methods such as phylogenetic tree analysis to infer the transmission and evolution of HIV [26, 155]. At the population scale, models of social and sexual contact networks are built to study the HIV epidemic [152, 105, 171, 145]. These models require estimation of many parameters such as frequency of sexual actions, infection probability per action, and parameters that shape the network structure. The validation of such models are based on data that have been collected from patients registered in hospitals or infected individuals who receive treatment.

In our previous study, Zarrabi et al 2012 [186] (Chapter 3), we tried

to address this issue by combining the data present at social and genetic scales in order to infer hypothetical transmission networks for HIV-1. The structure and properties of the inferred networks were in agreement with the recognized network structures for social and sexual contacts in the HIV-1 infected populations. We found that the transmission network of homosexual males and heterosexuals have a heterogeneous structure with a majority of "peripheral nodes" that have only a few sexual interactions and a minority of "hub nodes" that have many sexual interactions. A correlation was also detected between higher number of out-going links and longer untreated infection periods in HIV infected patients. These findings signified the importance of early treatment and supported the potential benefit of wide population screening, management of early diagnoses and anticipated antiretroviral treatment to prevent viral transmission and spread. The study was only considering the transmission within the three main risk groups of HIV: MSM, heterosexuals contacts, and injecting drug users (IDU). However, people belonging to different risk groups, as a part of one population, may have interactions with each other. For example, homosexuals males who have heterosexual contacts (bisexuals) or injecting drug users who have sexual activities [143, 102]. Therefore, the HIV infection can transmit from one risk group to the other.

In this chapter, we included mixing risk groups in the model in order to study transmission between different risk groups. For this, we modified the filter-reduction method presented in Chapter 3 to include mixture between risk groups. We analyzed the properties of inferred transmission networks and compared the results with those in Chapter 3. In what follows we briefly describe the proposed method for combining social and genetic data. Then we explain the modification of the method and use it to study transmission between risk groups in an Italian patient's database.

## 4.2 Combining Social and Genetic Networks

We proposed a new filter-reduction method for network construction. The method was used to convert raw and unstructured patient data

from social and genetic scales into network forms. Then we overlaid the networks from both scales and computed the intersection.

### 4.2.1 Filter-Reduction Method

The filter-reduction method was defined to convert raw patient data (i.e. social and clinical data and viral genetic sequence data) into network forms. The method was based on real patients data and no pre-assumptions were made on the network structure. The method was defined as follows: Let it be a graph/network composed by a set of $N$ nodes, $V(n)$. We started from an undirected fully-connected network of $V(n)$ in which there is a link between each pairs of nodes. A set of filters $F$ were applied on the fully-connected network which reduced the number of edges through the filtering process. Depending on the data and type of the network the filtering process may change. We built a social contact network of HIV-1 infected patients based on their social, clinical and demographical information. The contact network was defined as a graph with a set of nodes (infected individuals) and edges (possible social/sexual relationships). For building the network a set of social/sexual filters were applied. These filters were basic epidemiological criteria drawn from plausible assumptions in clinical studies such as, individuals belonging to the same age range or similar transmission risk group are more likely to have interactions. The effect of treatment in reducing the transmission probability was also considered in the filtering process.

In Table 3.1 we have presented the specific filtering rules that are used for building the social contact network. Age filter indicates the maximum age range for an individual to be socially or sexually interactive with another individual. Risk group filter considers the patient's gender ($g$) and risk group ($r$) and consists of three rules. First, the connection between patients from different risk groups is removed, which results in creation of three isolated clusters corresponding to the major HIV transmission risk groups (MSM, Heterosexual, and IDU). Second, for the heterosexual risk group the connection between patients with the same gender is removed. Third, the "Blood product" risk groups are isolated from the population, as they were not infected

through sexual relationships. Treatment filter considers an observation in clinical studies that the transmission probability of HIV-1 decreases by 80-98% after a patient starts treatment [37, 10]. The reason is mainly due to the smaller amount of viral particles in the genital secretions after treatment and the behavioural changes in the patients sexual and social habits when they become aware of their disease. Following this observation, we filtered connections to a patient $A$ from any other patient whose therapy initiation date ($t$) pre-dated patient $A$'s estimated seroconversion date ($s$). A direct connection between every two nodes that did not satisfy these epidemiological criteria was removed from the network. Then we inferred a meta-network of HIV-1 sequences based on the corresponding patient's demographic and medical information. We used the term "social contact network" as it contains all the contacts that are socially and sexually possible between infected individuals in the population. The inferred network is a directed acyclic graph (DAG) [169] that is a directed graph with no directed cycles. DAGs are suitable to study and model processes in which information flows in a consistent direction through the network such as disease transmission [60, 118]. Therefore, we used a DAG to model the transmission of HIV-1.

To test our method, we used a region-wide cohort study of HIV-1 infected people in Rome and Lazio region, Italy. A total of 655 patients were included in the analysis. The patients were annotated with demographics and genetic information, including: sequence id (numeric), viral subtype, sequence calendar year (numeric), patient's gender (male/female), age (numeric), mode of HIV transmission (MSM, heterosexual, IDU, blood products, other/unknown), country of origin (Italian/non-Italian/unknown), ART status (ART-experienced/ART-naive), seroconversion year (median time between last HIV-1 negative test date and first HIV-1 positive test date), calendar year of first HIV positive test and of first available antiretroviral therapy (numeric), plasma HIV-RNA load (numeric) at viral sequencing time. The missing numerical values were replaced with the average. These information were used to build the contact network by applying the filter-reduction method. To see the effect of applied filters we measured the fraction of removed edges by $E/E_{total}$, where $E$ is the number of edges

Table 4.1: Percentage of links filtered from the network by applying different filters.

| | Age filter | Risk group | Treatment | All filters |
|---|---|---|---|---|
| MSM | 44.5% | 0.0% | 69.2% | 80.4% |
| Heterosexual | 37.5% | 55.1% | 67.7% | 91.2% |
| IDU | 15.7% | 0.0% | 38.2% | 45.7% |
| All risk groups | 34.7% | 75.0% | 61.1% | 91.3% |

in the network after the filtering process and $E_{total}$ is the total number of edges in a directed fully-connected network. $E_{total}$ equals to $n(n-1)/2$, where $n$ is the number of nodes in the network.

Table 4.1 shows the percentage of links filtered from the network by applying each individual filter and all filters for different risk groups. By comparing the percentages for different risk groups, one can see that the age-filter removed more edges from the MSM and heterosexual networks than from the IDU. The risk group filter had the highest impact on all risk groups which results in three separate sub-networks corresponding to each risk group. The first rule of this filter separated the three risk groups and therefore confined the study to transmission within risk groups, omitting transmission between risk groups. The filter also caused the generation of a bipartite network for the heterosexual population. This was an effect of the second rule of the risk group filter, in which we considered two populations with different genders, males ($g_1$) and females ($g_2$), and only links between different genders were allowed. The effect of treatment filter on removing edges from the MSM and heterosexual was almost twice more than from the IDU.

For the phylogenetic analysis, the viral genomic RNA sequences of the patients in the same Italian cohort study were used. The sequence data information encompassed the HIV pol gene region (1-99 amino acids), covering the whole protease and most of the reverse transcriptase gene (at least the first 1-250 amino acids). HIV sequences were aligned using MUSCLE software [54] and the resulting multiple alignment was edited in order to remove drug-resistance associated mutations that can lead to convergent evolution bias in the phyloge-

netic tree estimation. A phylogenetic tree was then estimated using the maximum likelihood FastTree software [129], assessing node reliability via the built-in Shimodaira-Hasegawa test. Phylogenetic clusters were extracted from partitioning the leaves in the phylogenetic tree using the PhyloPart java application [130]. The PhyloPart uses a depth-first algorithm to extract a crisp partition (i.e. clustering) from an input phylogenetic tree, constraining its search on the comparison between sub-tree (i.e. potential clusters) and whole-tree patristic distance distributions, plus additional ancillary topologic criteria. When the sub-tree is highly (> 90%) supported by bootstrap (or posterior probability or other statistical tests), when at least two distinct patients are in the sub-tree, and when the median patristic distance is below a percentile threshold of the whole-tree distance distribution, then a cluster is found. If the depth-first search reaches a leaf node without finding any cluster, then the instance is classified as a singleton (A sequence that is not assigned to any cluster).

A total of 61 clusters (from size 2 to 52) are identified, where 39% of all patients are included in these clusters. A genetic distance matrix was also calculated with the MEGA software using the LogDet function [166]. The genetic distance matrix was used to build a genetic network. The distance matrix gave a weighted fully connected network which connected all sequences with each other using their genetic distances as weights. The connection between every two nodes with a genetic distance higher than a certain threshold was removed from the network. We used the threshold value of 0.04 nucleotide substitutions per site to build a genetic network. The threshold of 0.04 corresponds to the 15th percentile of the overall distance distribution measured through phylogenetic tree. The sense is that all retained links include sequences that are closer than the 85th percentile of the all pairwise comparisons (see [130] for a discussion on the optimal threshold). The fraction of removed edges from the genetic network using the threshold value 0.04 is 91.1% for the MSM, 77.5% for the heterosexual, 56.3% for the IDU and 76.5% for all risk groups. The percentage of removed edges from the IDU network is less in compare to the MSM and heterosexual networks which is consistent with the percentage of removed edges after applying the social filters.

### 4.2.2 Overlaying Networks

To combine information from both social and genetic scales we overlaid the networks from each scale and compute the intersection network. Two networks that are overlaid have exactly the same set of nodes. The edges in one network represent the social relationships and in the other represent the genetic connectivity. The intersection network contains only edges that are present in both networks. Figure 4.1 shows the workflow for converting unstructured data into network forms and computing their intersection. We compute the intersection of the social contact network with the genetic distance network. The socio-genetic intersection network is a hypothetical transmission network, which satisfies both social criteria (Table 3.1) and genetic criteria (genetic distance $< threshold$) for transmission events.

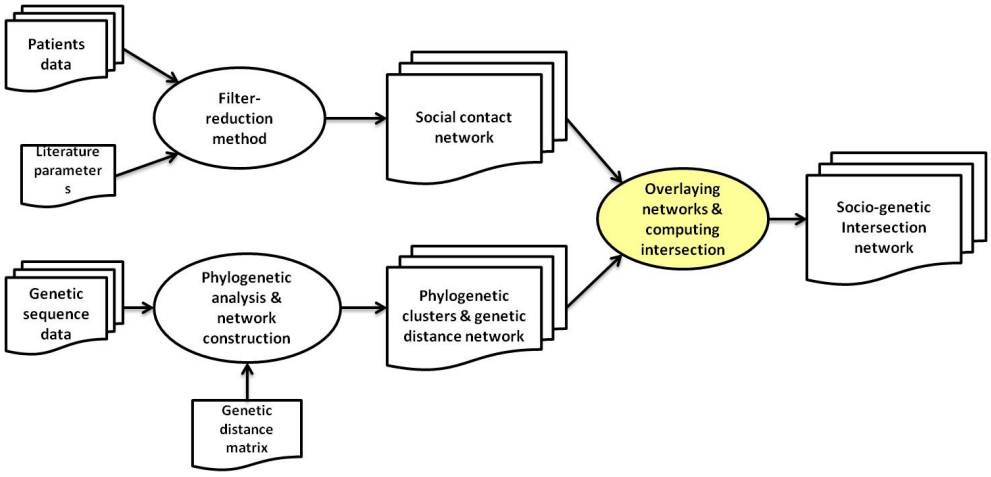


Figure 4.1: Overlaying social and genetic networks

For the MSM and heterosexual populations, the out-degree distributions were fitted to a straight line, in log-scale, with slopes equal to $2.65 \pm 0.43$ and $1.88 \pm 0.31$ Fitting to a straight line in a log-log scale suggested that the degree distribution follows a power-law with a scaling factor equal to the slope alpha [114, 40]. To ensure the fit to the power-law distribution we performed a statistical test, using maximum-likelihood fitting methods with goodness-of-fit tests based

on the Kolmogorov-Smirnov statistic as proposed by Newman et al. 2007 [40]. The power-law distribution for the MSM and heterosexual out-degree distributions yielded a scale-free structure for these networks with exponents equal to 2.65 and 1.88. This means that the structure of the hypothetical transmission network for the MSM and heterosexual population is heterogeneous, consisting of a majority of 'peripheral nodes' that have only a few sexual interactions and a minority of 'hub nodes' that have many sexual interactions. This finding was in line with the results obtained from analysis of the degree distribution of HIV transmission networks for the MSM population in the UK [29].

We further analysed other network properties such as fraction of removed edges, average degree, average path length, global and local clustering coefficients and anssortativity. Clustering coefficient gives an indication of the overall clustering in the network and is based on triplets of nodes [77]. A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected edges. A triangle consists of three closed triplets, one centred on each of the nodes. The global clustering coefficient is the ratio of the triangles and the connected triples in the graph. Local clustering coefficient of a node is the ratio of the triangles connected to the node and the triples centered on the node. The degree assortativity is a measure of connectivity between nodes with similar degrees [112]. The assortativity coefficient is positive if the connected nodes tend to have similar degrees and is negative otherwise. In Table 3.7 the network properties of intersection network in comparison to randomized networks is presented. We generated random networks of the same size (nodes and edges) as the computed intersection networks for each population (MSM, heterosexual, IDU and all risk groups). For this, we used the fraction of remaining edges in each network, as a probability to generate an edge in the random network. One can see that the inferred networks were different from random networks of their own size by having lower average path lengths, higher clustering coefficients and higher degree assrotativity coefficients.

## 4.3 Transmission Between Risk Groups

Transmission between different risk groups is important in HIV epidemic [128, 85, 82]. Individuals belonging to different risk groups of HIV may have contacts as part of a population. For example, homosexuals males who have heterosexual contacts (bisexuals) or injecting drug users who have sexual activities [143, 102]. Behaviorally bisexual men have long been known as a potential bridge for the entry of HIV into the heterosexual female population [128, 85]. In the United States, the percentage of homosexual males who are known to be bisexual is reported between 10% to 15% in 2001[39]. Sex partnerships with IDU is another risk factor that bridges the infection from IDU risk group to MSM and heterosexual populations. IDU partnership among New York city high-risk heterosexuals was reported to be 13.8% in 2006-2007 [82]. Thus, the study of transmission between risk groups is important in HIV epidemic in order to assess the relative role of bisexuality or IDU partnership in crossing the transmission of HIV between risk groups.

To include transmission between risk groups, we modified the second rule of the risk group filter in Table 3.1 which separates transmission risk groups. The new filtering rule keeps connections between patients based on phylogenetic clusters and certain probabilities. Phylogenetic clusters are a set of genetic clusters obtained through phylogenetic analysis (see Sect. 4.2). Nodes, representing individual viral isolates, residing in the same cluster are identified to be genetically close and therefore, possibly transmitted from one host to another. However, it is not known who has transmitted to whom as each cluster consist of a set of fully-connected nodes. A probability matrix is defined to specify the rates of sexual contacts between individuals from the same and different risk groups (Table 4.2). The modified risk group filter is shown in Algorithm 1, in which *cluster* represents the phylogenetic cluster, $r$ is the risk group, $p_{ij}$ is the rate of having contact between individuals from different risk groups, and $p_{ii}$ is the rate of having contact between individuals from the same risk group. If two individuals reside in the same phylogenetic cluster there is a connection between them in the network. Otherwise, a random number,

between zero and one, is generated as a probability to keep the connection based on the contact rates. In the probability matrix, we used the value of 0.7 for $p_{ii}$ and 0.1 for $p_{ij}$.These are free parameters in the model and are chosen as such to be close to the the values reported in the literature [82, 109, 120].

**if** $cluster_1 = cluster_2$ **then**
  $connection = 1$;
**else**
$rand \leftarrow (0 \leq$ random number $\leq$ 1);
  **if** $r_1 \neq r_2$ **then**
    **if** $rand \geq p_{ij}$ **then**
      $connection = 0$;
    **end if**
  **else**
    **if** $rand \geq p_{ii}$ **then**
      $connection = 0$;
    **end if**
  **end if**
**end if**

Algorithm 1: Modified risk group filter

| | MSM | HET | IDU |
|---|---|---|---|
| MSM | $p_{11}$ | $p_{12}$ | $p_{13}$ |
| HET | $p_{21}$ | $p_{22}$ | $p_{23}$ |
| IDU | $p_{31}$ | $p_{32}$ | $p_{33}$ |

Table 4.2: Risk-to-risk contact rates

We visualize the resulting networks using an in-house developed interactive visualization tool, called "Twilight"[1], which is based on the igraph software package for complex network research [43]. Twilight provides several algorithms to generate a layout of a network. We used a force-based layout algorithm developed by Fruchterman and Reingold [61]. This algorithm employs an iterative force-directed placement that considers the network as a mechanical system where the vertices are steel rings and the edges act as springs. Two basic principles govern the layout produced by this algorithm: (1) vertices connected by an edge should be drawn near each other, and (2) vertices should not be drawn too close to each other. The algorithm iteratively

[1]Twilight will be made available from http://uva.computationalscience.nl

seeks an equilibrium state that reduces the energy in the system from an initial random configuration to a configuration in which the energy state is minimal. Hence, the position of nodes in the network is a measure of their connectivity, meaning that nodes that are more connected to each other are also placed closer to each other in the network layout.

The hypothetical contact network is visualized in Figure 4.2.

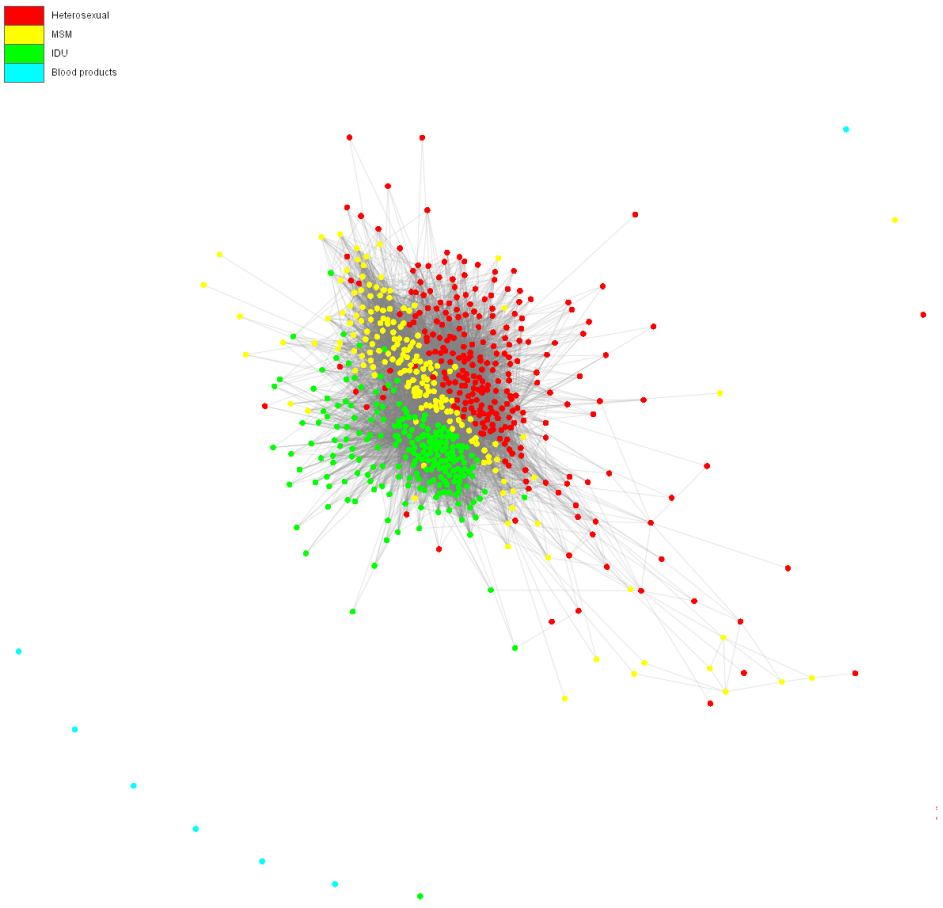


Figure 4.2: Hypothetical contact network.

Individuals are colored based on their transmission risk group: MSM (yellow), Heterosexual (red), and IDU (green). There were a few patients with blood products mode of infection (light blue) which were

isolated from other risk groups. The network in this case is a big connected component of most nodes and includes mixing between risk groups. It's clear through visualization that nodes from the same risk groups are more connected to each other than to nodes from other risk groups.

The intersection network is visualized in Figure 4.3 and the corresponding total-, in-, and out-degree distributions are shown in Figure 4.4.

This network contains fewer edges and is sparser than the contact network. Many isolated nodes exist in the network, which are the nodes that do not satisfy either the social or genetic criteria for transmission. From the visualization (Figure 4.3), one can see that the heterosexual individuals (red nodes) lie between the MSM and IDU. This suggests that, heterosexuals have tighter connections with both IDU and MSM, while the MSM and IDU are less connected. An interpretation could be that both bisexual behavior among Italian MSM and IDU partnership are relatively important in heterosexual transmission of HIV in Italy.

We measured properties of the inferred transmission network (with mixing risk groups) and compared that with properties of the inferred network in our previous study (with no mixing risk groups). The average degree in the case of mixing risk groups is almost half than with no mixing risk groups. The average path length is higher while, the clustering coefficients (local and global) and assortativity coefficients are lower. The data is presented in Table 4.3. Properties of the network with mixing risk groups is an average over the properties of 5 networks.

## 4.4 Conclusions and Future Directions

Considering population level data beside genomic data is essential for understanding the true nature of infectious disease transmission networks. The current research of HIV epidemic either uses genetic information of patients to infer the past infection events or uses statistics of sexual interactions to model the network structure of viral spreading.

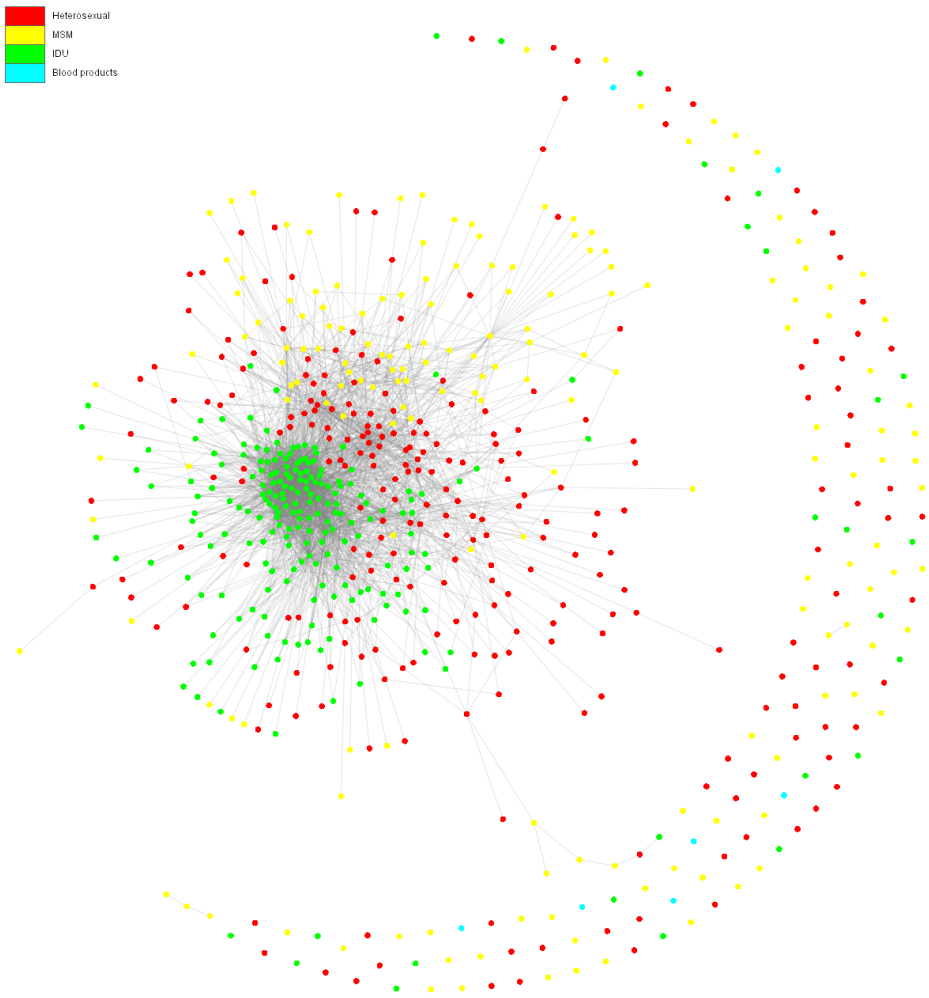


Figure 4.3: Hypothetical transmission network.

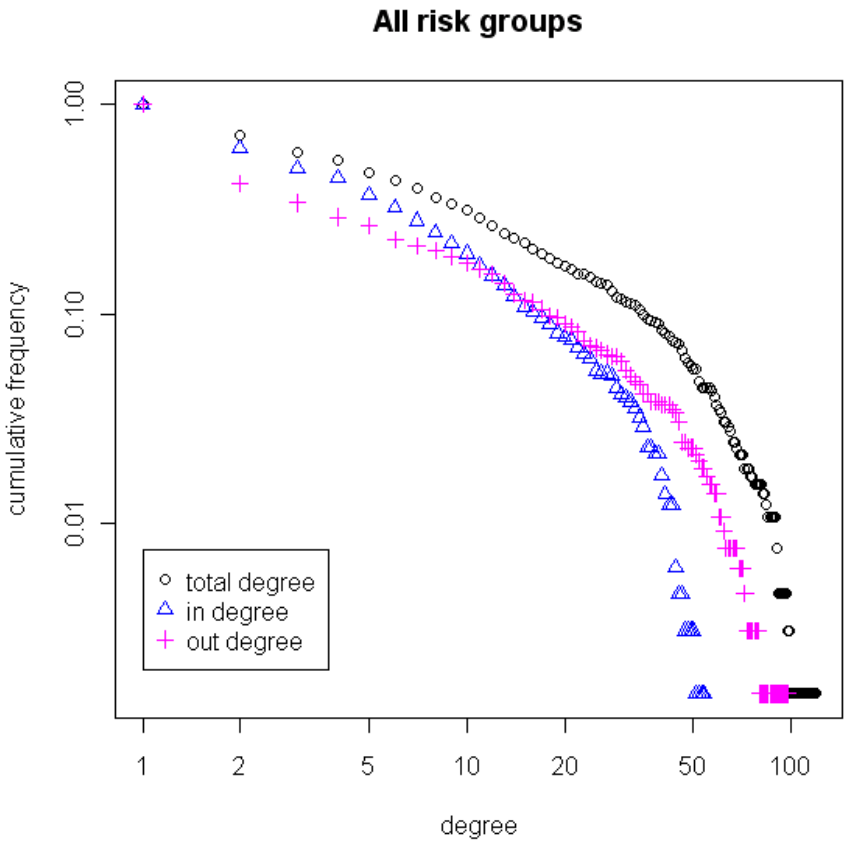


Figure 4.4: Total-, in-, and out-degree distributions of the hypothetical transmission network

Methods for a reliable reconstruction of HIV-1 transmission dynamics, taking into account both molecular and societal data, are still lacking. This issue was addressed in our previous study by proposing a new method to infer hypothetical HIV transmission networks by combining social and genetic networks. The method converts the data from both social and genetic scales into network forms and combines them by overlaying the networks and computing their intersection. The inferred networks suggested that the MSM and heterosexual population

Table 4.3: Properties of the hypothetical transmission network with and without mixing rirk groups.

| Networks properties | All risk groups | |
|---|---|---|
| | mixing risk groups | no-mixing |
| average degree | 12.39 | 21.10 |
| average path length | 2.71 | 2.22 |
| clustering coeff (global) | 0.25 | 0.59 |
| clustering coeff (local) | 0.26 | 0.45 |
| assortativity coeff (degree) | -0.03 | 0.11 |

has a heterogeneous structure with a majority of "peripheral nodes" that have only a few sexual interactions and a minority of "hub nodes" that have many sexual interactions. However, the study was limited to the transmission within the three major HIV risk groups (MSM, Heterosexual and IDU), omitting the transmission between risk groups. Infection transmission between different risk groups is important in HIV epidemic, as was also observed in the performed phylogenetic analysis [130]. In this study, we use the proposed method to infer networks of HIV infected patients by considering mixture among major HIV risk groups. We modified the filtering process in the filter-reduction method in order to include mixing risk groups in the model. For this, we use the information on phylogenetic clusters obtained through phylogenetic analysis. A probability matrix is also defined to specify contact rates between individuals form the same and different risk groups. We apply this method to reconstruct networks of HIV infected patients in central Italy. Our results suggests that bisexual behavior among Italian MSM and IDU partnership are relatively important in heterosexual transmission of HIV in central Italy.