VIRUS EVOLUTION

Virus Evolution, 2023, **9(1)**, 1–6
DOI: https://doi.org/10.1093/ve/vead026
Advance access publication date: 25 April 2023
Reflections

From components to communities: bringing network science to clustering for molecular epidemiology

Molly Liu,[1] Connor Chato,[1] and Art F. Y. Poon[1,2,3,†,*]

[1]Department of Pathology and Laboratory Medicine, Western University, Dental Sciences Building, Rm. 4044, London, ON N6A 5C1, Canada, [2]Department of Microbiology and Immunology, Western University, 1151 Richmond Street, London, ON N6A 3K7, Canada and [3]Department of Computer Science, Western University, Room 355, Middlesex College, London, ON N6A 5B7, Canada
[†]https://orcid.org/0000-0003-3779-154X
*Corresponding author: E-mail: apoon42@uwo.ca

Abstract

Defining clusters of epidemiologically related infections is a common problem in the surveillance of infectious disease. A popular method for generating clusters is pairwise distance clustering, which assigns pairs of sequences to the same cluster if their genetic distance falls below some threshold. The result is often represented as a network or graph of nodes. A connected component is a set of interconnected nodes in a graph that are not connected to any other node. The prevailing approach to pairwise clustering is to map clusters to the connected components of the graph on a one-to-one basis. We propose that this definition of clusters is unnecessarily rigid. For instance, the connected components can collapse into one cluster by the addition of a single sequence that bridges nodes in the respective components. Moreover, the distance thresholds typically used for viruses like HIV-1 tend to exclude a large proportion of new sequences, making it difficult to train models for predicting cluster growth. These issues may be resolved by revisiting how we define clusters from genetic distances. Community detection is a promising class of clustering methods from the field of network science. A community is a set of nodes that are more densely inter-connected relative to the number of their connections to external nodes. Thus, a connected component may be partitioned into two or more communities. Here we describe community detection methods in the context of genetic clustering for epidemiology, demonstrate how a popular method (Markov clustering) enables us to resolve variation in transmission rates within a giant connected component of HIV-1 sequences, and identify current challenges and directions for further work.

1. Introduction

Identifying groups of closely related infections is a common problem in epidemiology. The distribution of infections in space or time is often used as proxy for their epidemiological relationships. In other words, infections that were sampled in a similar location, at a similar time, or both, may share a common source. The genetic similarity of infections can be a more convenient or informative proxy than space or time, particularly for infections that can establish a persistent, chronic infection; that can remain undiagnosed as an asymptomatic infection; and/or with a relatively low rate of transmission. For instance, there is an abundance of genetic clustering studies characterizing patterns of transmission of HIV-1 (Grabowski and Redd 2014) and hepatitis C virus (Lamoury et al. 2015). Moreover, genetic sequences are often routinely collected as a part of public health surveillance and the clinical management of infections.

There is now an extensive literature on the use of 'molecular' or 'genetic' clusters to characterize patterns of transmission in a population (Poon 2016; Hassan et al. 2017). Clustering on the basis of the pairwise distances among sequences (Balfe et al. 1990; Aldous et al. 2012), as measured by a genetic distance (d) is especially popular in part because these distances can be relatively fast to compute. Moreover, pairwise distances are immutable quantities; unlike phylogenies, they do not change with the addition of sequences to the database. Any pair of sequences that have a distance below some threshold are assigned to the same cluster. We can describe this process more formally as follows: consider a complete graph $G = (V, E)$, where each vertex $v \in V$ represents a sequence or an individual infection. Every edge $e(v,u) \in E$ between vertices $v, u \in V$ is weighted by the genetic distance between the respective sequences, $d(v,u)$. Applying a distance threshold $d_{\max}$ yields a subgraph of G that retains the full set of vertices and a reduced set of edges, $G' = (V, E')$, where $E' = \{e(v,u) \in E : d(v,u) \leq d_{\max}\}$.

A connected component is a maximal subgraph $G_c = (V_c, E_c)$ of G such that any vertex $v \in V_c$ can be reached from any other vertex $u \in V_c$ through a path of edges in $E_c \subseteq E$. Any given G_c cannot be contained within a larger connected component. Although it is seldom stated explicitly, studies that use pairwise genetic clustering almost always define clusters as connected components of at least two or more vertices. Thus, even though a single vertex is considered a component in graph theory, it is generally not interpreted as a cluster of size one in the context of infectious disease. Indeed, these 'non-clustered' vertices are often excluded

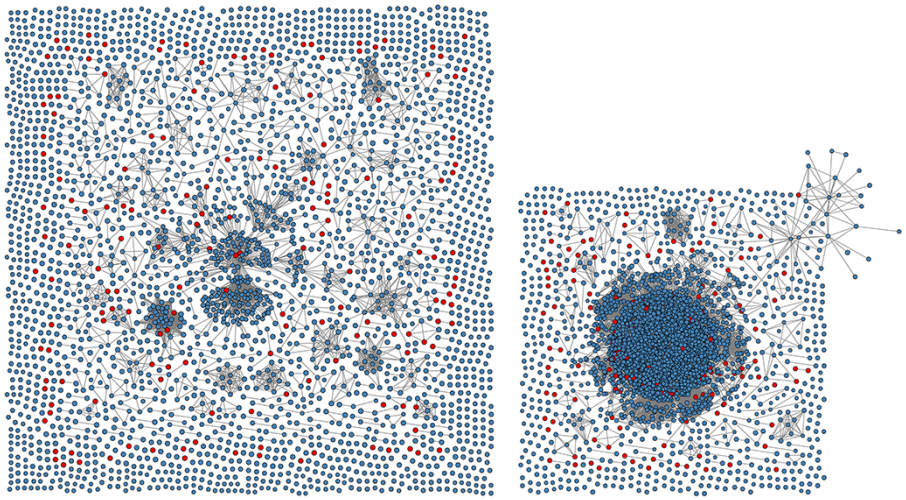

Figure 1. Visualizations of the graphs generated by applying thresholds of $d_{\max} = 0.015$ (left) and $d_{\max} = 0.03$ (right) to the pairwise distance matrix for $n = 2,915$ HIV-1 sequences from Dennis et al. (2018). The graph layouts were generated using the 'neato' algorithm in GraphViz (Ellson et al. 2001). Each point represents an HIV-1 infection, with its area scaled in proportion to its year of sampling. Points are colored red (non-blue) if the infection was sampled in the most recent year of the study (2015), and blue otherwise.

from visualizations of the connected components. This separation of vertices into clustered and non-clustered categories is frequently used as a surrogate binary variable to assess potential transmission risk factors through logistic regression (Aldous et al. 2012; Poon et al. 2015; Ragonnet-Cronin et al. 2019). The size and composition of the connected components are determined by the distance threshold. With increasing values of $d_{\max}$, the vertices gradually coalesce into one giant connected component. Conversely, as $d_{\max}$ approaches zero, each vertex becomes isolated into its own component. Thus, clustering studies employ intermediate thresholds that yield a number of connected components of moderate size. This also tends to result in a substantial number of unclustered vertices.

The cross-sectional and prospective analyses of genetic clusters are a rapidly developing area of molecular epidemiology. For instance, several studies have developed models to predict the addition of newly diagnosed people to pre-existing clusters in a population database (Ragonnet-Cronin et al. 2016a; Wertheim et al. 2018; Bachmann et al. 2020). The ability to predict where the next cases will appear in the population would have tangible public health applications (Billock et al. 2019), providing more timely and actionable information than the retrospective characterization of cluster growth in the past. It also provides a statistical basis for optimizing $d_{\max}$ to given population (Chato, Kalish and Poon 2020). In our previous work, however, we also observed that a substantial fraction ($> 50\%$) of sequences representing new diagnoses did not become connected to any clusters at typical distance thresholds, making them impossible to predict.

Our postulate is that the conventional practice of defining clusters from connected components is a limiting and unnecessary constraint on this predictive application of molecular epidemiology. Specifically, there are several studies in network science that have developed algorithms that can further partition connected components into smaller clusters (Fortunato 2010; Leskovec et al. 2009; Karrer and Newman 2011). These are known as community detection methods. For example, the Louvain algorithm (Blondel et al. 2008) employs a 'bottom-up' heuristic to search for the assignment of vertices to clusters that maximizes the modularity of the graph. Modularity is a statistic that compares the observed number of edges within clusters to a random graph (Newman 2006). Community detection methods are predominantly associated with the analysis of large social networks (Bedi and Sharma 2016), particularly in relation to social media (Papadopoulos et al. 2012). However, they have also been applied to biological clustering problems, *e.g.*, predicting protein function from sequence homology (Enright, Van Dongen and Ouzounis 2002), protein interaction networks (Gulbahce and Lehmann 2008), and gene expression networks (Treviño III et al. 2012). In sum, this abundant literature on community detection represents an untapped resource for improving applications of genetic clustering for infectious disease epidemiology.

2. Example application to HIV-1 sequences

To demonstrate the use of community detection for genetic clustering, we obtained 2,915 anonymized HIV-1 *pol* sequences from GenBank (accession numbers MH352627–MH355541). These sequences were used in a retrospective study of HIV-1 transmission patterns among people attending the Vanderbilt Comprehensive Care Clinic in middle Tennessee, USA (Dennis et al. 2018). We generated a multiple sequence alignment using MAFFT version 7.3.10 (Katoh and Standley 2013) and used the program TN93 (https://github.com/veg/tn93) to calculate the pairwise genetic distances using the Tamura and Nei (1993) formula. The resulting graphs at thresholds of $d_{\max} = 0.015$ and 0.03 are displayed in Fig. 1. At the 1.5 per cent threshold used in the original study, only 65 (40.1 per cent) of 162 'new' nodes sampled in the last year of the study were connected to clusters of sequences sampled prior to 2015. Increasing the threshold to 3.0 per cent augments this number from 62 to 118 (72.8 per cent). However, this also causes the graph to coalesce around a giant connected component of 1,752 nodes. The number of connected components of size 2 or greater decreases from 253 to 73.

Next, we used the Poisson regression method that we previously developed (Chato, Kalish and Poon 2020) to determine the optimal $d_{\max}$ threshold for these data. The underlying concept

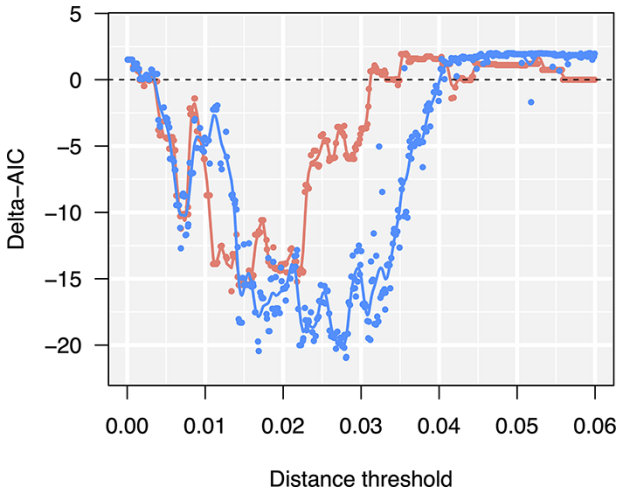

Figure 2. ΔAIC profiles for connected component (red/non-blue) and Markov clustering (MCL, blue) methods under a range of Tamura–Nei (TN93) distance thresholds. More negative ΔAIC values indicate less information loss when incorporating additional predictor variables into a Poisson regression of new nodes among clusters (Chato, Kalish and Poon 2020). Each point represents one of 420 parameter combinations, specifically the distance threshold ($d_{\max}$) and the expansion (k) and inflation (r) parameters of the MCL method. Solid lines correspond to cubic smoothing splines fit to each set of points.

is that the optimal threshold should yield a distribution of new nodes among connected components (as clusters) that we can predict the most accurately, based on measurable characteristics of those clusters. This is quantified by the difference in the Akaike information criteria (AIC) of two Poisson regression models. The null model uses only the size of a cluster as a predictor variable, which is equivalent to assuming that every infection has the same probability of being the most closely related to a new infection (at a distance below $d_{\max}$). An alternate model incorporates additional predictor variables, in this case the mean time since sampling for nodes in a cluster (Chato, Kalish and Poon 2020). We calculated the AIC of both models under a range of thresholds to yield a profile. In short, the ΔAIC for connected components was minimized at $d_{\max} = 0.0134$ (Fig. 2), which was fairly similar to the threshold used in the original study (0.015).

Finally, we applied a community detection method known as the Markov cluster algorithm (Van Dongen 2008) to the graphs obtained under varying thresholds, using the implementation of this method in the R package MCL (https://CRAN.R-project.org/package=MCL). MCL acts on a transition matrix (P) derived from the graph. In our case, we start from the adjacency matrix (A) of the undirected graph, where: $A_{ij} = 1$ if there exists an edge between vertices i and j, and 0 otherwise; $A_{ii} = 0$; and $A_{ij} = A_{ji}\ \forall\ i \neq j$. To derive P from A, we normalize the entries so that each column sums to 1, i.e., $P_{ij} = A_{ij} / \sum_k A_{kj}$. Next, two different matrix operations are iteratively applied to P. The inflation operation takes the r^{th} Hadamard (entry-wise) power of P, such that $(P_{ij})^r = P_{ij}^r$ and then rescales the result so that its columns each sum to 1. The expansion operation takes the k^{th} power of P by matrix multiplication; for example, $P^k = PP$ for $k = 2$. These operations are analogous to simulating a random diffusion process through the graph (Van Dongen 2008). This iterative algorithm is applied until it converges to an equilibrium state where the matrices before and after operations are identical, or up to a maximum number of iterations.

We used Latin hypercube sampling to generate a uniform sample of 500 points in the space of all three parameters over the respective continuous ranges: $0 \leq d_{\max} \leq 0.6$; $2 \leq k \leq 25$; and $2 \leq r \leq 25$. Out of these MCL analyses, 80 (16 per cent) failed to converge to an equilibrium matrix after 100 iterations. These failures tended to be associated with $d_{\max} < 0.025$ or $d_{\max} > 0.04$. We repeated the Poisson regression analysis on clusters produced by the MCL method to generate a ΔAIC profile with respect to $d_{\max}$ (Fig. 2). To minimize the effect of varying k and r on estimating the optimal distance threshold, we located the minimum of a cubic smoothed spline fit to these ΔAIC values, resulting in $d_{\max} = 0.0276$. This turned out to be very close to the threshold associated with the parameter combination with the lowest ΔAIC, $d_{\max} = 0.028$.

The most conspicuous effect of MCL is that it partitions the largest connected component, which comprises 1,860 sequences at $d_{\max} = 0.028$, into 403 clusters (Fig. 3A). At this threshold, the largest component grows by 73 new nodes. These nodes become redistributed among 25 (6.2 per cent) of the clusters (Fig. 3B). We can also see that the clusters within this largest component that accumulated one or more new nodes in 2015 tended to have more recent sampling dates than inactive clusters of the same size. Thus, even though the majority of nodes have become subsumed into a single giant component, we are still able to resolve the epidemiological variation among clusters of nodes within this component. Furthermore, these effects of cluster size and mean sampling dates on the distribution of cluster growth within the largest connected component can be shown to be dependent on $d_{\max}$ (Supplementary Fig. S2).

3. Challenges and future directions

The use of connected components has become so routine for interpreting the graphs defined by the pairwise distances among virus sequences that the term 'clusters' have become synonymous with connected components. Community detection methods provide a useful extension of the connected components approach because 'giant' components can be broken down into more informative clusters. This confers greater scalability; for instance, clusters are sometimes used as foci for computationally intensive analyses (*e.g.*, Lewis et al. 2008). As we have demonstrated above, community detection also enables the user to relax the distance threshold and thereby capture a larger proportion of observed cluster 'growth' for analysis. Otherwise, an excessive number of new sequences become excluded from training models for forecasting cluster growth.

One basic challenge to incorporating community detection methods into the molecular epidemiology toolkit is that there are numerous and diverse methods to choose from. In addition to MCL and the Louvain algorithm, for instance, there is also stochastic blockmodeling (Karrer and Newman 2011), convolutional neural networks (Jin et al. 2021a), and methods based on random fields (He et al. 2018); see Jin et al. (2021b) for a recent review. This may engender confusion in the field of molecular epidemiology, where many different genetic clustering methods have already been developed (e.g., McCloskey and Poon 2017; Villandré et al. 2018; Han et al. 2019; Volz et al. 2020). Some public health agencies have already committed to a specific method of genetic clustering, such as the US Centers for Disease Control and Prevention and HIV-TRACE (Oster et al. 2018; Kosakovsky Pond et al. 2018). Thus, there would doubtless need to be some demonstrable superiority of community detection over the *status quo* for these methods to see application in the public health domain. In addition, none of these community detection methods were

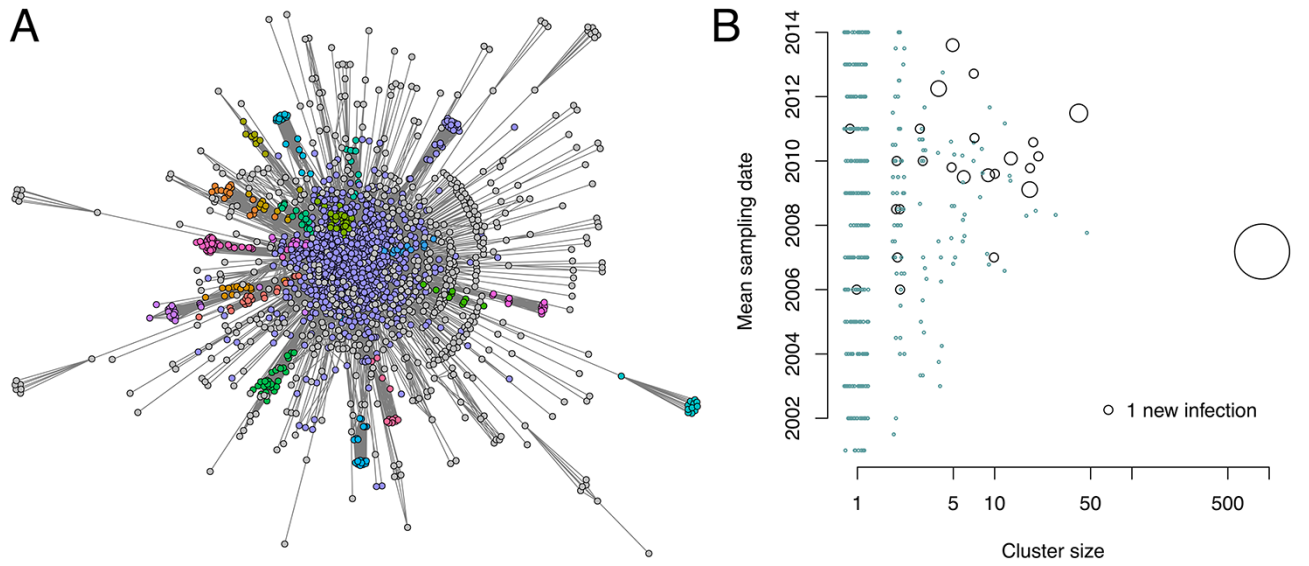

Figure 3. (A) Visualization of the largest connected component of HIV-1 sequences when $d_{\max} = 0.028$. The vertices are colored with respect to the 20 largest clusters as determined by the MCL algorithm and gray otherwise. Unlike Fig. 1, we used the scaleable force directed placement algorithm (*sfdp* in GraphViz; Hu 2005) to generate a layout of this subgraph that emphasizes the separation of clusters. A more color-accessible version with varying node shapes and a different layout algorithm is provided as Supplementary Fig. S1. (B) Bubble plot summarizing the number of new cases among MCL clusters in the largest connected component. Each point represents a cluster, with its area scaled in proportion to the number of new nodes added to the cluster in 2015. The smallest points, drawn in blue, represent clusters with zero new nodes. We added a random 'jitter' to cluster size to reduce overlap.

designed specifically for infectious disease epidemiology. Indeed we have not found any example in the literature of such methods being used to characterize viral transmission dynamics by clustering genetic sequences—at best, there is limited prior work for potential users to reference.

Another potential challenge of applying community detection to genetic clustering studies is that the resulting clusters may be unstable to the addition of new data. One of the useful features of connected components derived from pairwise distances is that they can only increase in size; it is not possible for a connected component to decrease in size with additional data. On the other hand, the addition of nodes to a connected component may change how a community detection method partitions the component into clusters (network communities). The number of clusters within the component may even increase or decrease as a result. However, this problem is not exclusive to community detection methods. Connected components can become merged by the addition of one or more sequences that fall between the two components, i.e., within the distance threshold to members of both components. Even a single new edge between two clusters is sufficient to merge them into a single component. In contrast, community detection methods should be more robust to the addition of these intermediate nodes, since edges between the communities will remain relatively sparse. Clusters that are derived from phylogenies, i.e. subtree clustering methods, are also not robust to the addition of sequences (Chato et al. 2022). Nevertheless, characterizing the sensitivity of community detection methods to the addition of data in the context of molecular epidemiology will be an important area for research.

Incorporating community detection to a clustering analysis can introduce more parameters to be calibrated by the user, in addition to distance or phylogenetic bootstrap thresholds (Hassan et al. 2017). The MCL method, for example, adds two parameters for the matrix inflation (r) and expansion (k) operations, respectively. However, we found that the ΔAIC profile that we used to optimize the distance threshold was relatively insensitive to variation in r and k (Fig. S3). These results suggest that our ability to predict the distribution of new infections may be more robust to differences in community detection methods, although we have only evaluated a small number of such methods in this context. In addition, community detection methods may be too computationally complex to apply to large sequence data sets. For instance, it is not uncommon to use genetic clustering to analyze a population database comprising tens of thousands of HIV-1 sequences or more (Poon et al. 2015; Ragonnet-Cronin et al. 2016b). Fortunately, community detection methods are often designed to handle very large networks (Harenberg et al. 2014), and some have already been adapted to distributed computing environments (e.g., Azad et al. 2018).

Genetic clustering can play an important role in tracking variation in virus transmission rates in near real-time (Little et al. 2014; Poon et al. 2016). However, this emerging practice of 'molecular surveillance' has also raised significant concerns over ethics, consent, and data privacy (Coltart et al. 2018). This is especially controversial for HIV-1, which remains a highly stigmatized infectious disease where people are criminally prosecuted for virus transmission. In this context, the phrase 'community detection' may be problematic, since it can be misinterpreted as an act of surveillance targeting actual communities. In many settings, communities are an important source of support, information and advocacy for people living with HIV-1 (Campbell, Nair and Maimane 2007). When communicating findings from applications of these methods to infectious disease epidemiology, we recommend making it clear that while community detection methods were largely developed for the analysis of social networks, they are being applied to networks where connections represent levels of genetic similarity between infections—not social links. Although networks are being used in both contexts, they are abstractions of completely different sets of relationships.

Community detection methods may be especially well-suited for pathogens with a higher transmission rate than HIV-1, such as SARS-CoV-2. When the rate of transmission exceeds the rate of molecular evolution, there is a low probability that an infection transmitted to the next host will have accumulated one or more mutations. Consequently, the distribution of pairwise distances will be shifted towards zero. In the case of SARS-CoV-2 genome sequences, setting the pairwise distance threshold to the equivalent of two nucleotide substitutions (about 6.7×10^{-5} expected substitutions per site) or more tends to result in giant connected components. Even at the lowest possible threshold of one mutation, we have found that pairwise distance clustering of SARS-CoV-2 genome sequences tends to yield enormous, densely connected components, making it difficult to identify associations between individual- and group-level characteristics and transmission patterns. Thus, community detection may provide an important mechanism enabling investigators to resolve transmission patterns from genetic sequences for a much wider range of viruses than HIV-1 and hepatitis C virus.

Data availability

Anonymized HIV-1 sequences from Dennis et al. (2018) are publicly available at GenBank (accession numbers MH352627–MH355541). Pre-processed data and the Python/R scripts used to generate all figures are publicly available on Zenodo at https://doi.org/10.5281/zenodo.7020457.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Funding

Canadian Institutes of Health Research (PJT-156178 and PJT-183832) to A.F.Y.P.

Author contributions

The study was conceived and designed by A.F.Y.P. M.L. and A.F.Y.P. curated the data. M.L. implemented the statistical and computational methods to analyze the data. C.C. provided analytical tools. A.F.Y.P. and M.L. drafted the manuscript and generated data visualizations. A.F.Y.P., M.L. and C.C. critically reviewed and revised the manuscript.

References

Aldous, J. L. et al. (2012) 'Characterizing HIV transmission networks across the United States', *Clinical Infectious Diseases*, 55: 1135–1143.

Azad, A. et al. (2018) 'HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks', *Nucleic Acids Research*, 46: e33–e33.

Bachmann, N. et al. (2020) 'Phylogenetic cluster analysis identifies virological and behavioral drivers of HIV transmission in msm', *Clinical Infectious Diseases*, 72: 2175–2183.

Balfe, P. et al. (1990) 'Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations', *Journal of Virology*, 64: 6221–6233.

Bedi, P. and Sharma, C. (2016) 'Community detection in social networks', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6: 115–135.

Billock, R. M. et al. (1999) 'Prediction of HIV transmission cluster growth with statewide surveillance data', *Journal of Acquired Immune Deficiency syndromes*, 80: 152.

Blondel, V. D. et al. (2008) 'Fast unfolding of communities in large networks', *Journal of Statistical Mechanics: Theory and Experiment*, 2008: 10008.

Campbell, C., Nair, Y. and Maimane, S. (2007) 'Building contexts that support effective community responses to HIV/AIDS: a South African case study', *American Journal of Community Psychology*, 39: 347–363.

Chato, C. et al. (2022) 'Optimized phylogenetic clustering of HIV-1 sequence data for public health applications', *PLOS Computational Biology*, 18: e1010745.

Chato, C., Kalish, M. L. and Poon, A. F. (2020) 'Public health in genetic spaces: a statistical framework to optimize cluster-based outbreak detection', *Virus Evolution*, 6: veaa011.

Coltart, C. E. et al. (2018) 'Ethical considerations in global HIV phylogenetic research', *The Lancet HIV*, 5: e656–e666.

Dennis, A. M. et al. (2018) 'HIV-1 transmission clustering and phylodynamics highlight the important role of young men who have sex with men', *AIDS Research and Human Retroviruses*, 34: 879–888.

Ellson, J. et al. (2001) In *International Symposium on Graph Drawing*. Graphviz—open source graph drawing tools, Springer, pp. 483–484.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002) 'An efficient algorithm for large-scale detection of protein families', *Nucleic Acids Research*, 30: 1575–1584.

Fortunato, S. (2010) 'Community detection in graphs', *Physics Reports*, 486: 75–174.

Grabowski, M. K. and Redd, A. D. (2014) 'Molecular tools for studying HIV transmission in sexual networks', *Current Opinion in HIV and AIDS*, 9: 126.

Gulbahce, N. and Lehmann, S. (2008) 'The art of community detection', *BioEssays*, 30: 934–938.

Han, A. X. et al. (2019) 'Inferring putative transmission clusters with Phydelity', *Virus Evolution*, 5: vez039.

Harenberg, S. et al. (2014) 'Community detection in large-scale networks: a survey and empirical evaluation', *Wiley Interdisciplinary Reviews: Computational Statistics*, 6: 426–439.

Hassan, A. S. et al. (2017) 'Defining HIV-1 transmission clusters based on sequence data', *AIDS (London, England)*, 31: 1211.

He, D. et al. (2018) 'A network-specific Markov random field approach to community detection', in *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hu, Y. (2005) 'Efficient, high-quality force-directed graph drawing', *Mathematica Journal*, 10: 37–71.

Jin, D. et al. (2021a) 'Heterogeneous graph neural network via attribute completion', in *Proceedings of the Web Conference 2021*, pp. 391–400.

Jin, D. et al. (2021b) 'A survey of community detection approaches: From statistical modeling to deep learning', in *IEEE Transactions on Knowledge and Data Engineering*.

Karrer, B. and Newman, M. E. (2011) 'Stochastic blockmodels and community structure in networks', *Physical Review E*, 83: 016107.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular Biology and Evolution*, 30: 772–780.

Kosakovsky Pond, S. L. et al. (2018) 'HIV-TRACE (TRAnsmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens', *Molecular Biology and Evolution*, 35: 1812–1819.

Lamoury, F. M. et al. (2015) 'The influence of hepatitis c virus genetic region on phylogenetic clustering analysis', *PloS one*, 10: e0131437.

Leskovec, J. et al. (2009) 'Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters', *Internet Mathematics*, 6: 29–123.

Lewis, F. et al. (2008) 'Episodic sexual transmission of HIV revealed by molecular phylodynamics', *PLoS medicine*, 5: e50.

Little, S. J. et al. (2014) 'Using HIV networks to inform real time prevention interventions', *PLoS ONE*, 9: e98443.

McCloskey, R. M. and Poon, A. F. (2017) 'A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation', *PLoS Computational biology*, 13: e1005868.

Newman, M. E. (2006) 'Modularity and community structure in networks', *Proceedings of the National Academy of Sciences*, 103: 8577–8582.

Oster, A. M. et al. (1999) 'Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data', *Journal of Acquired Immune Deficiency Syndromes*, 79: 543.

Papadopoulos, S. et al. (2012) 'Community detection in social media', *Data Mining and Knowledge Discovery*, 24: 515–554.

Poon, A. F. (2016) 'Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks', *Virus Evolution*, 2: vew031.

Poon, A. F. et al. (2015) 'The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada', *The Journal of Infectious Diseases*, 211: 926–935.

Poon, A. F. et al. (2016) 'Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study', *The Lancet HIV*, 3: e231–e238.

Ragonnet-Cronin, M. et al. (2016a) 'Transmission of non-B HIV subtypes in the United Kingdom is increasingly driven by large non-heterosexual transmission clusters', *The Journal of Infectious Diseases*, 213: 1410–1418.

Ragonnet-Cronin, M. L. et al. (2016b) 'A direct comparison of two densely sampled HIV epidemics: the UK and Switzerland', *Scientific Reports*, 6: 1–9.

Ragonnet-Cronin, M. et al. (2019) 'HIV transmission networks among transgender women in Los Angeles County, CA, USA: a phylogenetic analysis of surveillance data', *The Lancet HIV*, 6: e164–e172.

Tamura, K. and Nei, M. (1993) 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees', *Molecular Biology and Evolution*, 10: 512–526.

Treviño III, S. et al. (2012) 'Robust detection of hierarchical communities from Escherichia coli gene expression data', *PLoS Computational Biology*, 8: e1002391.

Van Dongen, S. (2008) 'Graph clustering via a discrete uncoupling process', *SIAM Journal on Matrix Analysis and Applications*, 30: 121–141.

Villandré, L. et al. (2018) 'DM-PhyClus: a Bayesian phylogenetic algorithm for infectious disease transmission cluster inference', *BMC Bioinformatics*, 19: 1–16.

Volz, E. M. et al. (2020) 'Identification of hidden population structure in time-scaled phylogenies', *Systematic Biology*, 69: 884–896.

Wertheim, J. O. et al. (2018) 'Growth of HIV-1 molecular transmission clusters in New York City', *The Journal of Infectious Diseases*, 218: 1943–1953.