AUTO-TUNE: FINDING AN OPTIMAL DISTANCE THRESHOLD FOR INFERRING HIV TRANSMISSION CLUSTERS

Steven Weaver [1*], **Vanessa Davila Conn** [2], **Hannah Verdonk** [1], **Joel Wertheim** [3], and **Sergei L. Kosakovsky Pond** [1]

[1] *Center for Viral Evolution, Temple University, Philadelphia, PA, USA*
[2] *Center for Research in Infectious Diseases, National Institute of Respiratory Diseases, Mexico City, Mexico*
[3] *Department of Medicine, University of California, San Diego, CA*

Correspondence*:
Steven Weaver
sweaver@temple.edu

ABSTRACT

Choosing an appropriate distance threshold is an important part of inferring a transmission network to determine the relative growth of clusters within a localized epidemic. This distance threshold determines how close two consensus sequences must be in order for a link to be created between them in the network. Using a distance threshold that is too high can result in a network with many unnecessary links, making it difficult to interpret and analyze. On the other hand, using a distance threshold that is too low can result in a network with too few links, which may not capture key insights into rapidly growing clusters among patients with shared attributes that could benefit from public health intervention measures.

Here, we present a heuristic scoring approach for tuning a distance threshold by associating each tested threshold against the maximal number of clusters created across all thresholds and the difference between the ratio (R_{12}) of the largest cluster in the network to the second largest cluster at each iteration. The number of clusters is normalized between $[0, 1]$ then gated via a Gompertz function transform. Meanwhile, the distribution of all R_{12} ratios are converted to Z scores, and normalized relative to the largest positive Z score across all candidate distances. The priority score is the sum of aforementioned two components.

Published research using the HIV-TRACE software package frequently use the default threshold of 1.5% for HIV pol gene sequences. We apply our scoring heuristic to outbreaks with different characteristics, such as regional or temporal variability, and demonstrate the utility of using the scoring mechanism's suggested distance threshold to identify clusters exhibiting risk factors that would have otherwise been more difficult to identify. For example, while we found that a 1.5% distance threshold is typical for US-like epidemics, recent outbreaks like the CRF07_BC subtype among men who have sex with men (MSM) in China has been found to have a lower optimal threshold of 0.5% to better capture the transition from injected drug use (IDU) to MSM as the primary risk factor. Alternatively, in communities surrounding Lake Victoria, where there has been sustained transmission for several years, we found that a larger distance threshold is suitable to

28 capture a more risk factor diverse populace with sparse sampling over a longer period of time.
29 Such identification may allow for more informed intervention action by respective public health
30 officials.

31 **Keywords: molecular epidemiology, HIV, network, transmission cluster, surveillance**

1 INTRODUCTION

32 Choosing an appropriate distance threshold is an important part of using a transmission network to track
33 the spread of a contagious disease. This distance threshold determines how close two individuals must be
34 in order for a link to be created between them in the network.

35 Using a distance threshold that is too small can result in a network with many unnecessary links, making
36 it difficult to interpret and analyze. On the other hand, using a distance threshold that is too large can result
37 in a network with too few links, making it difficult to accurately track the spread of the disease.

38 To ensure that the transmission network is useful and informative, it is important to carefully consider the
39 appropriate distance threshold. This may vary depending on the specific disease and the context in which it
40 is spreading. For example, a highly contagious respiratory illness may require a smaller distance threshold
41 than a less contagious illness that is primarily spread through direct contact.

42 In general, the goal is to strike a balance between having enough links to accurately track the spread of
43 the disease, while not having so many links that the network becomes difficult to interpret. This can be
44 achieved through careful analysis and consideration of the specific disease and context.

45 Overall, choosing an appropriate distance threshold is an important step in using a transmission network
46 to track the spread of a contagious disease. It can help ensure that the network is useful and informative,
47 and can ultimately aid in efforts to control and prevent the spread of the disease.

2 METHODS

2.1 Scoring Heuristic Procedure

49 Network threshold selection procedure proceeds as follows:

50 1. For each candidate threshold d_L, in increasing order, ranging from the smallest genetic distance in
51   the dataset, up to either the largest distance or a predetermined maximal threshold, we compute two
52   network statistics: R_{12}, the ratio of the largest cluster to the second largest cluster, and C – the number
53   of clusters in the network.

54 2. A priority score is assigned to each d_L. This score measures two properties of the threshold: Does R_{12}
55   jump at d_L? How far is the number of clusters C at d_L from the maximal number of clusters over all
56   threshold values? Let there be N overall d_L candidate values, and assume we are examining the ith
57   candidate, d_L^i with $W < i \leq N - W$ (W is a positive integer defined below).

58   a. The R_{12} jump is computed by looking at the normalized ratio of the mean R_{12} values computed over
59   the leading window $d_L^{i+1} \ldots d_L^{i+W}$ and the trailing window $d_L^{i-W} \ldots d_L^{i-1}$. The width of the window,
60   W, is defined as $(([\frac{N}{100}], 3), 30)$. The distribution of ratios is converted to Z scores, and normalized
61   relative to the largest positive Z score across all candidate distances, yielding the jump component
62   of the score.

63   b. The number of clusters, C_i at threshold d_L^i is first normalized to $[0, 1]$ through $\frac{C_{max}-C_i}{C_{max}-C_{min}}$ and next

64   gated via a Gompertz function transform $1 - e^{-e^{-25x+3}}$ This function provides an ad hoc means for

65   penalizing having too few clusters relative to the maximum over all ranges. For example, a threshold

66   that yields 95% of the maximal number of clusters receives a score of 0.996, while a threshold that

67   yields 85% - a score of 0.376.

68   c. The priority score for d_L^i is the sum of the two components defined in (a) and (b).

69 3. The threshold with the highest priority score will be selected as the suggested automatic distance

70   threshold, if the score is high enough (1.9 or more), and either of the two conditions hold.

71   a. No other thresholds have priority scores of 1.9 or higher

72   b. If other thresholds have priority scores of 1.9 or higher, then the range of thresholds represented by

73   these options is small (no more than $log(N)$ times the mean step between successive d_L^i).

74 4. If no single threshold can be selected in step 3, then the one with the highest priority score is suggested,

75   and an inspection of the plot like the one on the analyze page is recommended to ensure that the

76   threshold is sensible.

2.2 Assortativity

78   Degree-weighted homophily (DWH) is a measure of similarity between nodes in a network based on

79 their attributes (such as demographic characteristics or behaviors) and their degree (i.e., the number of

80 connections they have to other nodes in the network). It is used to quantify the extent to which nodes with

81 similar attributes tend to be connected to each other more frequently than would be expected by chance.

82 DWH is calculated as the ratio of the observed number of connections between nodes with similar attributes

83 to the expected number of connections between such nodes, based on their degree.

84   In mathematical terms, it is defined as:

$$DWH = \frac{W_M + W_C - 2W_X}{\frac{d_{in}}{nodes_{in}^2} + \frac{d_{out}}{nodes_{out}^2}} \tag{1}$$

85 Where

86 • W_M : Weight of in-group connections

87 • W_C : Weight of out-group connections

88 • W_X : Weight of cross-group connections

89 • d_{in} : In-group degree

90 • d_{out} : Out-group degree

91 • $nodes_{in}$: number of in-group nodes

92 • $nodes_{out}$: number of out-group nodes

93   DWH ranges from -1 to 1. A DWH value of 0 indicates that there is no more homophily than expected

94 with chance, while a value of 1 indicates that there is perfect homophily (e.g. Birds always link to birds,

95 and only birds). A value of -1 is achieved for perfectly disassortative networks (e.g. Bird never linking with

96 another bird).

97 DWH is used in social network analysis and in the study of how different attributes are related to the
98 formation of connections between individuals. It is used as a way to measure the similarity of attributes
99 between individuals in a network.

2.3 Implementation

101 The software implementation involves a step-by-step process that utilizes the HIV-TRACE suite of
102 packages. It starts with calculating pairwise distances wth the tn93 tool and a supplied multiple sequence
103 alignment. This generated pairwise distances are supplied to the hivnetworkcsv script while providing the
104 -A keyword argument. A brief outline of the software's implementation are as follows

105 1. Calculate pairwise distances: The user first calculates the pairwise distances using the tn93 fast pairwise
106   distance calculator, providing the necessary threshold value and the input FASTA file. The command
107   for this step is

```
1  tn93 -t 0.030 pol.fasta > pairwise_distances.15.tn93.csv
```

111 Please note that the threshold should include the maximal range one is intending to test.

112 2. Compute distance threshold scores: The hivnetworkcsv script is then executed with the required input
113   file, format, and autotune option to generate a tab-separated output file, as shown below

```
1  hivnetworkcsv -i pairwise_distances.15.tn93.csv -f plain -A 0 > autotune_report.tsv
```

117 3. Visualize the report: Users can upload the generated autotune_report.tsv file to
118   `http://autotune.datamonkey.org/analyze` for visualization and further analysis of the
119   data. This web-based platform provides an interactive environment to explore scores and other metrics
120   across the range of tested outputs.

121 4. Run HIV-TRACE: Once AUTO-TUNEd threshold(s) are settled upon after review, the user runs the
122   HIV-TRACE command with the appropriate input FASTA file, distance threshold, and other required
123   arguments. The output is saved as a JSON file. An example command is

```
1  hivtrace -i ./INPUT.FASTA -a resolve -r HXB2_prrt -t < autotune_threshold > -m 500 -g .05
   ↪> hivtrace.results.json
```

2.3.1 Optional : Compute Assortatviity Metrics

129 5. Annotate results: The hivnetworkannotate script is used to annotate the results obtained from the
130   HIV-TRACE step with attributes. The script takes the JSON results file, node attributes file, schema
131   file, and a resolve flag as input.

```
1  hivnetworkannotate -n hivtrace.results.json -a node_attributes.json -g schema.json -r
```

135 For more information, users can refer to the hivnetworkannotate documentation.

136 6. Analyze the results with DWH: After the results file has been annotated, the user can proceed to the
137   assortativity page, http://autotune.datamonkey.org/assortativity, for further analysis of the output.

138 AUTO-TUNE is readily accessible on GitHub as part of the hivclustering repository
139 (https://github.com/veg/hivclustering). It is integrated into the command-line interface of the software
140 as the -A or –auto-profile argument. hivclustering is a key component of the HIV-TRACE suite of
141 tools, a resource for the inference, analysis and visualization of HIV transmission networks.

142 The Degree Weighted Homophily (DWH) calculation tool, an integral component of the assortativity
143 step, is developed using TypeScript, a statically typed superset of JavaScript that ensures robustness
144 and scalability. In an effort to promote accessibility and ease of integration, the DWH tool is packaged
145 and distributed through the Node Package Manager (NPM), enabling researchers and developers to
146 conveniently incorporate this advanced analytical tool into their own projects and workflows. DWH
147 can be used in-browser or as a command line tool, allowing researchers and developers to employ
148 the tool in an interactive command-line interface or integrate it into larger software applications, thus
149 catering to a diverse array of technical needs and preferences. Instructions for usage and installation is
150 found on Github (https://github.com/veg/dwh).

151 The described workflow offers a systematic approach to analyze potential distance thresholds for one's
152 data with AUTO-TUNE, from calculating pairwise distances to visualizing and annotating results.

2.4 Visualization

154 Visualizations of AUTO-TUNE results are accessible at http://autotune.datamonkey.org/analyze. It is a
155 dynamic and interactive web-based platform that offers visualization and analysis of results generated by
156 AUTO-TUNE. The website provides a comprehensive view of the data by generating various plots across
157 candidate distance thresholds. These include a score plot, allowing users to identify trends and anomalies
158 across the full range of thresholds. Additionally, it generates a graph showing the number of clusters across
159 candidate thresholds, one of the components that contribute to the score. The site also includes an R1/R2
160 plot that displays the ratio of the largest cluster to the second largest cluster across candidate thresholds,
161 which is the other metric that contributes to the scoring heuristic.

162 An assortativity tool is available at http://autotune.datamonkey.org/assortativity, and is an advanced
163 analytical tool engineered to facilitate the calculation of Degree Weighted Homophily (DWH) values. It
164 utilizes the DWH NPM package to generate a tabular representation of DWH values corresponding to each
165 value for a selected attribute annotation, providing an exhaustive examination of the interrelationships for
166 the field. A notable feature is the computation of the panmictic range, which involves a label permutation
167 test to generate the null distribution of DWH values. This feature establishes a comparative baseline that
168 aids in determining the significance of homophily versus what would be expected by chance. Lastly, the site
169 also provides a plot of the fraction of pairwise connections, normalized by degree, for each value pertinent
170 to the selected field. This visual depiction facilitates an intuitive comprehension of the distribution and
171 interconnections within the dataset.

172 The site aims to offer a user-friendly interface for data visualization, playing an important role in
173 interpreting and understanding AUTO-TUNE's output data. The visualization code is available on Github
174 (https://github.com/stevenweaver/autotune-app/).

2.5 Data curation and analysis

3 RESULTS

3.1 Case Study 1: Middle Tennessee, Seattle, and Alberta

3.2 Case Study 2: Comparison with Prior Publications

3.3 Case Study 3: Comparison with Prior Publications

4 DISCUSSION

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

The Author Contributions section is mandatory for all articles, including articles by sole authors. If an appropriate statement is not provided on submission, a standard one will be inserted during the production process. The Author Contributions statement must describe the contributions of individual authors referred to by their initials and, in doing so, all authors agree to be accountable for the content of the work. Please see here for full authorship criteria.

FUNDING

Details of all funding sources should be provided, including grant numbers if applicable. Please ensure to add all necessary funding information, as after publication this is no longer possible.

ACKNOWLEDGMENTS

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder.

DATA AVAILABILITY STATEMENT

The datasets [GENERATED/ANALYZED] for this study can be found in the [NAME OF REPOSITORY] [LINK].

TABLES

Table 1. Homophily Comparison

Dataset	Art's Threshold Avg. Homophily	AUTO-TUNE Avg. Homophily
TN	0.0079	0.0147
Seattle	0.0259	0.0348
Canada	-0.0536	-0.0448

5 FIGURE CAPTIONS

Table 2. Paper Comparison

PMID	Country	Collection Date	Threshold Used	AUTO-TUNE
29975689	South Africa	2011-2015	2.5%	2.584%
30574123	USA	1997-2008	2%	1.848%
32500089	China	2008-2015	0.5%	0.675%
32693608	Uganda	2009-2016	1.5%	1.707%
33281803	China	2000-2016	0.5%/0.7%	0.676%
33901684	China	2008-2012	1.5%	1.215%
34452506	Canada	1996-2017	1.5%/2.5%	0.547%
31041344	USA	1997-2017	1.5%	0.927%