**All text: Times New Roman 12-pt. font; double-spaced and left aligned (left-justified).**

**Manuscript Number** (if known)

**Article Summary Line:** Graph databases provide a computational method to integrate community surveillance, phylogenetic, and epidemic networks into a single platform for the identification of transmission clusters.

**Running Title:** Metadata and Phylogeny Graph Database Proof of Concept

**Keywords:** Database Management Systems, Phylogenetic Relationships, Molecular Epidemiology

**Abstract**

To illustrate a proof of concept for a graph database integrating molecular sequence data, metadata, and community surveillance data, we segmented 22,713 SARS-CoV-2 samples into 5 groups, generated maximum likelihood trees, and derived a inferred transmission network from a forest of minimum spanning trees built on samples’ patristic distances.

**Text—word count** (word counts are strictly enforced)

Integration of phylogenetic, epidemiologic, and contact tracing networks provides insights into transmission dynamics and aids in cluster identification, but this approach is still a nascent field. Many tools available to integrate these data for a single analysis such as MicrobeTrace and StrainHub focus on providing a visualization tool for end users and are limited by local hardware.1,2 COVID-19 provides a unique challenge due to the sheer volume of both sequencing data and metadata. We need an infrastructural approach to integrate these data to ensure they are tractable in the long-term. We propose a graph database as a source of record to integrate these data and identify transmission clusters at scale. Graph databases, a type of NoSQL databases, differ from relational database management systems (RDBMS) in that data are stored as nodes, relationships, and properties rather than a series of tables joined together. The primary benefit from such a design is that computation involving primarily relationships between entities are easily traversed, making network computations more efficient through indexes along these paths.3

The Study

We obtained metadata from 22,713 samples collected August-September 2021 in Texas and downloaded full sequences from GISAID for each sample. We aligned sequences to the Wuhan-Hu-1 reference genome using MAFFT 7.313, then split the dataset into 5 subsets to mimic an ongoing pandemic response with biweekly updates: 08/01/2021-08/14/2021, 08/01/2021-08/28/2021, 08/01/2021-09/11/2021, 08/01/2021-09/25/2021, 08/01/2021-09/30/2021. For each subset, we generated a maximum likelihood (ML) tree under a generalized time-reversible model of nucleotide evolution using FastTree 2.1.11.

To investigate the viability of leveraging a graph database integrating community surveillance and contact tracing with phylogenetic network analysis, we designed a schema connecting our ML trees with sample metadata which we loaded into an instance of Neo4j Desktop 1.4.3, a graph database which uses cypher query language. We added trees to the database as tree alignment graphs (TAG), an approach developed as part of the Open Tree of Life Project, allowing us to share metadata across multiple trees with potentially conflicting relationships. This approach also provides the flexibility to generate trees using subsets or supersets of existing trees.4 Using our 5 ML trees, we calculated the patristic distance between each sample for each tree, which provided a simulation of periodic updates to a phylogenetic tree as samples are collected. While pairwise genetic distance is oft used in transmission clustering due to its low computational cost, patristic distances use more of the information available allowing for easier differentiation between rapidly and slowly evolving sites.5

From the patristic distance networks, we then generated a forest of 100 (out of 463 possible) minimum spanning trees (MST), each starting from a different initial node collected August 1, 2021. While distance thresholds can be used to infer transmission, MSTs are less prone to discarding links. We then aggregated the resultant edges (as counts and average patristic distances) to build our inferred transmission network. However, molecular sequence data alone is insufficient to truly infer transmission events given the rapid transmission of SARS-CoV-2. In the future, we would benefit from leveraging these networks to connect disjoint contact tracing and venue affiliation data.

We then used Neo4j Bloom to create visualizations for each of the 5 aggregate networks. For each of the 5 data subsets, we generated a graph of samples connected by their aggregated MST network with edge weights determined by percentage of MSTs that share the edge. We used node color to show clusters’ relationship with metadata elements. As one might expect, Pango lineage was strongly associated with individual clusters (Figure 1), but (when available) age and race did not appear associated with clusters (Figure 2). While the demographic fields we leveraged failed to overwhelmingly identify within specific clusters upon visual inspection, we need to further investigate through a regression model if there exists a discernible pattern between them.

Conclusions

As we intend for this work to be a proof of concept, there are several limitations we did not address. In addition to the lack of contact tracing and venue affiliation data to corroborate the distance-based transmission estimates, we did not include reference sequences from neighboring regions, preventing us from identifying quantity or origin of variant introductions. While limited in the inferences we can draw from this network, in the future, we can integrate this network with contact tracing and venue affiliation data, using molecular sequence data to fill in the gaps in our known transmission networks. On a computational note, we were able to render networks of this scale in Neo4j Bloom (an exploratory data tool), but we started to run into its limits with just over 125,000 nodes and relationships. Future work should include more light-weight, static visualization scripts for common use cases or interactive visualizations without active physics engines.

While this initial analysis could be generated without necessitating a graph database, we benefit at scale with such an approach. Graph databases frequently include millions, and in some applications, trillions, of nodes and relationships.6,7 Many network analysis algorithms, e.g. community detection, similarity, pathfinding, and node embedding, have implementations available in Neo4j.8 Incorporating additional data layers such as contact tracing and venue affiliation data should enable for more accurate inference of putative transmission networks. A more immediate next step would be to leverage clustering techniques such as k-means or HDBSCAN to separate cluster signal from noise and more precisely derive clusters.9

We developed a Python helper script to parse Newick tree files and several Cypher scripts (Neo4j graph database query language) to populate the database, calculate the patristic distances, and generate MSTs across the graph. <https://github.com/glstott/bahllab>

**Acknowledgments**

**Disclaimers**

A disclaimer is not generally needed because EID includes a [disclaimer in online and in print.](http://wwwnc.cdc.gov/eid/page/about)

**Author Bio** (first author only, unless there are only 2 authors)

In 2–3 sentences, include author’s current position, affiliation, and primary research interests.

**Footnotes (if applicable)**

**References**

Do not use endnotes for references. References should appear in a numbered list, in the order in which they are cited in the manuscript, including citations in figures and tables. Format the list in hanging indent paragraph style. Use [Uniform Requirements citation style](https://www.ncbi.nlm.nih.gov/books/NBK7282/) for authors and publications. List first 6 authors followed by “et al.” Reference limits are strictly enforced.

1. Campbell, E. M., Boyles, A., Shankar, A., Kim, J., Knyazev, S., Cintron, R., & Switzer, W. M. (2021). MicrobeTrace: Retooling molecular epidemiology for rapid public health response. *PLoS computational biology*, *17*(9), e1009300. <https://doi.org/10.1371/journal.pcbi.1009300>
2. de Bernardi Schneider, A., Ford, C. T., Hostager, R., Williams, J., Cioce, M., Çatalyürek, Ü. V., Wertheim, J. O., & Janies, D. (2020). StrainHub: a phylogenetic tool to construct pathogen transmission networks. *Bioinformatics (Oxford, England)*, *36*(3), 945–947. <https://doi.org/10.1093/bioinformatics/btz646>
3. Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., & Wilkins, D. (2010, April). A comparison of a graph database and a relational database: a data provenance perspective. In *Proceedings of the 48th annual Southeast regional conference* (pp. 1-6).
4. Smith, S. A., Brown, J. W., & Hinchliff, C. E. (2013). Analyzing and synthesizing phylogenies using tree alignment graphs. *PLoS computational biology*, *9*(9), e1003223. <https://doi.org/10.1371/journal.pcbi.1003223>
5. Ellsworth M Campbell, Hongwei Jia, Anupama Shankar, Debra Hanson, Wei Luo, Silvina Masciotra, et al., Detailed Transmission Network Analysis of a Large Opiate-Driven Outbreak of HIV Infection in the United States, *The Journal of Infectious Diseases*, Volume 216, Issue 9, 1 November 2017, Pages 1053–1062, <https://doi.org/10.1093/infdis/jix307>
6. Ching, A., Edunov, S., Kabiljo, M., Logothetis, D., & Muthukrishnan, S. (2015). One trillion edges: Graph processing at facebook-scale. *Proceedings of the VLDB Endowment*, *8*(12), 1804-1815.
7. Have, C. T., & Jensen, L. J. (2013). Are graph databases ready for bioinformatics?. *Bioinformatics (Oxford, England)*, *29*(24), 3107–3108. <https://doi.org/10.1093/bioinformatics/btt549>
8. Needham, M., & Hodler, A. E. (2018). A comprehensive guide to graph algorithms in neo4j. *Neo4j. com*.
9. McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205.

A picture containing background pattern

Description automatically generated

Figure 1. Inferred transmission networks colored by race and age for the entire 2 month timeframe. In both panels, grey indicates no response. In the race panel colors correspond to the following: red indicates Asian, orange for black or African American, brown for Native HI or Pacific Islander, yellow for white, green for multiple races, blue for other race. In the age panel, the increasing hue of purple corresponds to increasing age.

Background pattern

Description automatically generated with medium confidence

Figure 2. Inferred transmission networks by time slice (selected three datasets for brevity), colored by Pango Lineage.