Red de virus de vertebrados a través del Property Graph Model

Mikaela Lezcano y Felipe Maresca

Resumen

El estudio del viroma global es de interés para la salud humana, animal, y la comprensión de los procesos ecológicos y evolutivos en general. En este trabajo usamos el Property-Graph Model para migrar la base de datos VIRION, que contiene registros de interacciones entre virus y vertebrados, y realizamos consultas para comprender si estas interacciones tienden a agruparse según una señal taxonómica, para así discernir si a lo largo de la evolución las interacciones entre virus y vertebrados han tendido a mantenerse o a diferenciarse. Para esto nos valemos de las ventajas que presenta representar los datos en una base de datos en grafos. Encontramos que para la resolución taxonómica de la clase, existe una señal de la taxonomía en la conformación de las comunidades de la red, pudiendo así afirmar que existe una conservación de nicho en las interacciones virus-vertebrado a lo largo de la evolución. El trabajo realizado se puede reproducir a través del código e instrucciones presentes en el siguiente repositorio https://github.com/FelipeMaresca/TrabajoFinalBDNR2023

I. Introducción

En el presente trabajo se estudiarán las relaciones entre animales vertebrados y virus de distinto tipo, dado que el viroma global, es decir el conjunto de todos los virus en toda la biosfera, es uno de los campos menos documentados de la biodiversidad global. Se estima que hay al menos 40.000 especies de virus que infectan solamente a los mamíferos, de los cuales miles probablemente pueden infectar a los humanos. Además existen millones de especies de virus más, cuyas interacciones se distribuyen a lo largo del árbol de la vida, pero sólo unos pocos pueden infectar a hospedadores de varias ramas con una profunda división evolutiva (Carlson et al., 2022). Al mismo tiempo, el estudio de estos datos, constituye una pieza fundamental para la evaluación del riesgo zoonótico (enfermedad transmitida entre animales no humanos que infecta a humanos), ya que los virus capaces de dar saltos amplios de hospedador son los más predispuestos a futuras emergencias.

Para ello, primero se migrará la red de virus de animales del esquema relacional al no relacional utilizando los modelos de base de datos de grafos, más específicamente el Property Graph Model (PGM). Luego se realizará una breve caracterización de la red a partir de algunas medidas de topología de la red y centralidad. Posteriormente, se aplicarán diferentes métodos de detección de comunidades en redes con la finalidad de indagar cuáles son los grupos de especies de animales y/o de virus más propensos a vincularse entre sí y si estos grupos están estructurados por su taxonomía. Para estudiar esto último, en primer lugar se realizará un test de hipótesis con las frecuencias de clases observadas y las esperadas por azar en las comunidades detectadas de mayor tamaño, y en segundo lugar se seguirá un abordaje de modelos nulos para detectar clases dentro de las comunidades con sobrerrepresentación y subrepresentación respecto al azar.

II. Objetivos e Hipótesis

II-A. Objetivos

El objetivo general del trabajo consiste en aplicar técnicas de bases de datos no relacionales a partir de una base de datos relacional con la intención de abordar el problema de las relaciones entre virus y vertebrados con una perspectiva “coast to coast”. Para esto se plantean un conjunto de objetivos específicos:

- Diseñar y migrar la base de datos relacional a una no relacional siguiendo el Property Graph Model (PGM).
- Analizar si existen patrones en las interacciones a nivel taxonómico, es decir determinar si existe sobredispersión o agrupamiento de las interacciones entre las distintas clasificaciones jerárquicas de virus y vertebrados.
- Reconocer subgrupos o comunidades de animales y virus fuertemente cohesionados a partir de la aplicación de alguna técnica de clústers/conglomerados para redes.
- Identificar grupos o especies de virus que representan o podrían representar un riesgo para la salud humana.

II-B. Hipótesis

A fin de cumplir los objetivos mencionados, se plantean las siguientes interrogantes: ¿qué clase de animales y de virus tienen más vínculos? ¿Los virus de un mismo grupo tienden a infectar a los mismos grupos de vertebrados (debido a posibles restricciones evolutivas) o tienden a diferenciarse en los animales que infectan (evolucionando para evitar la competencia entre ellos)?

Hipótesis: Las interacciones de virus-hospedador podrían, a priori, mostrar dos tendencias opuestas, una al agrupamiento a niveles taxonómicos y otra opuesta, a la sobredispersión. Debido a restricciones evolutivas, es probable que los atributos que permiten la infección de un cierto grupo sean conservadas, y por ende virus emparentados entre sí tiendan a infectar a los

mismos grupos, ya que el desarrollo de una maquinaria infecciosa es un proceso complejo. Por otra parte, es probable que virus muy emparentados hayan tenido que desarrollar alguna clase de diferenciación en el uso de los recursos, por lo tanto infectando a animales diferentes. El balance entre las fuerzas opuestas de diferenciación de nicho (diferenciación en el uso de los recursos), y la conservación de nicho (restricciones a los posibles usos de los recursos presentadas por la evolución) determinará qué tanto se diferencian los virus de un mismo grupo en sus interacciones. La existencia de hipótesis plausibles contrapuestas hace necesario un análisis de los datos que indique cuál de las dos es la correcta.

III. Los Datos

Para llevar a cabo el trabajo se utilizó una base de datos de asociaciones de virus vertebrados, llamada The Global Virome in One Network (VIRION) (Carlson, 2022). La misma es mantenida dinámicamente y se ha construído a partir de la agregación de un total de cinco fuentes estáticas (denominadas HP3, GMPD2, EID2, Shaw y PREDICT) y dos fuentes dinámicas (Global Biotic Interactions y GenBank, hospedadas por el Centro Nacional de Información Biotecnológica [NCBI]).[1]

Todas estas fuentes de datos están armonizadas en una columna vertebral taxonómica, incluidos los metadatos sobre la validez taxonómica del huésped y el virus y la clasificación superior; metadatos adicionales sobre la metodología de muestreo y la fuerza de la evidencia también están disponibles en un formato armonizado.

Se trata de la base de código y acceso abierto más grande disponible, con aproximadamente medio millón de registros únicos que incluyen 9.521 “especies“ de virus resueltas, 3.692 especies de huéspedes vertebrados resueltas y 23.147 interacciones únicas entre organismos taxonómicamente válidos. Juntos, estos datos cubren aproximadamente una cuarta parte (1635 especies) de la diversidad de mamíferos, una décima parte (1072 especies) de la diversidad de aves y el 6 % de la diversidad total estimada de vertebrados (60.000), alcanzando una proporción mucho mayor de su viroma que cualquier base de datos anterior Este es un ejemplo de texto con una nota al pie de página.[2] . La base de datos VIRION consta de 6 tablas (ver Figura 2): vínculos, taxonomía del huésped o animal (especie, género, familia, orden, clase), taxonomía del virus (especie, género, familia, orden, clase), procedencia (de que base de datos proviene la información del vínculo), detección (forma en que fue detectado el vínculo) y temporal (información temporal respecto a la detección del mismo, por ejemplo año de la publicación). En Anexos se puede ver la definición de cada uno de los campos.

Figura 1. Esquema conceptual de la base de datos relacional VIRION

[1]Hoy en día, la base de datos se actualiza automáticamente mediante GitHub Actions. Todas las noches, las fuentes dinámicas (GenBank y GLOBI) se vuelven a descargar, se someten a una reconciliación taxonómica completa y se reincorporan a la base de datos. La actualización también se puede ejecutar manualmente en su totalidad para actualizar la taxonomía en las fuentes estáticas (CLOVER y PREDICT), produciendo una nueva compilación estable. Repositorio de GitHub etiquetado por versión: https://github.com/viralemergence/virion.

[2]Se debe tener en cuenta que cada uno de estos grupos tiene sus propios sesgos y brechas de cobertura únicos en términos de taxonomía del huésped y muestreo geográfico.

IV. METODOLOGÍA

IV-A. Modelos de base de datos de grafos - Property Graph Model (PGM)

Los modelos de base de datos de grafos proporcionan una excepcional herramienta para estudiar relaciones entre diferentes actores. Al mismo tiempo, se consideran especialmente útiles cuando las entidades están muy conectadas a través de relaciones descriptivas y además existen relaciones de muchos a muchos entre las entidades, como es en este caso. Actualmente, uno de los sistemas más populares a implementar dentro de estos es el Modelo Property Graph (PGM por sus siglas en inglés). El mismo consiste en pseudografos dirigidos y parejas clave-valor llamadas propiedades las cuales están asociadas a nodos y aristas. Además es posible etiquetar tanto los nodos como las aristas con más de una etiqueta.

Para implementarlo, lo primero a decidir es cómo representar los elementos de la realidad en términos de nodos, etiquetas (conjunto de nodos con determinada característica), relaciones y propiedades (atributos o características). Las propiedades son importantes porque permiten almacenar datos sobre las relaciones y los nodos. En este caso los nodos junto con sus etiquetas serán los nodos tipo Host y los nodos tipo Virus, y la relación entre los mismos estará dada a partir de la etiqueta "Infecta a", entendiendo que los Virus infectan a los Host y por lo tanto la dirección del vínculo va en ese sentido, pero también podría haberse definido en el sentido contrario con la etiqueta "Infectado por". Todas las características de Host y Virus serán propiedades de los nodos pero las características asociadas a la relación, serán propiedades del vínculo. El grafo resultante constituye una red bipartita, por lo que además se realizará la proyección de misma, en la que los nodos Host están unidos por un vínculo "Shares_V" si tienen al menos 1 virus en común, y de forma análoga estarán vinculados entre sí los nodos Virus con relaciones "Shares_H" (ver Figura 1).

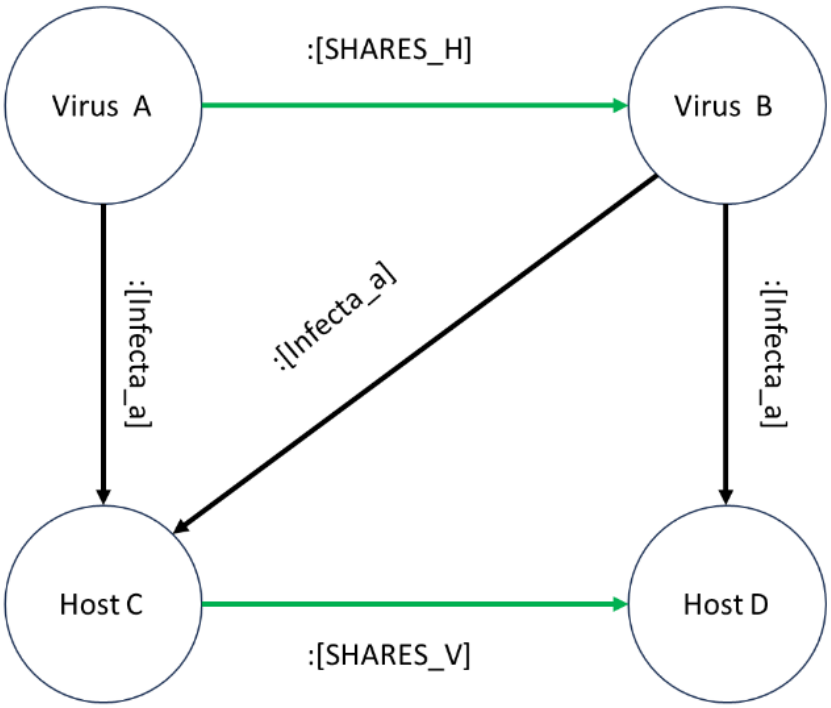

Figura 2. Esquema representativo de la red general y las redes proyectadas. Los vínculos en negro son aquellos que fueron creados usando la información presente en la base de datos original y conforman la red bipartita. Usando la información allí contenida, se realizaron las proyecciones de esta red para los nodos tipo Host y tipo Virus. Si dos hospedadores tienen vínculos con el mismo virus, entonces fueron unidos por un vínculo "SHARES_V", y de forma análoga para los virus se crearon los vínculos "SHARES_H".

Modelar este problema en particular como una base de datos en grafos presenta varias ventajas. En primer lugar, la flexibilidad asociada a estos modelos de datos permite representar como nodos a los actores (virus y animales), pero también utilizar esta representación para los grupos (familias, clases, órdenes, etc) lo cual permite realizar las consultas pertinentes a este problema de forma más simple. Además, la representación a través de un grafo o red permite visualizar mejor el problema, dado que por su propia naturaleza implica relaciones del tipo N:N entre varios actores de distintas categorías. La representación mediante un modelo de Property Graph también presenta ventajas para realizar los análisis pertinentes de estructura de la red. El proyecto presentado, fue implementado en Neo4j, "software libre de Base de datos orientada a grafos, implementado en Java. Los desarrolladores describen a Neo4j como un motor de persistencia embebido, basado en disco, implementado en Java, completamente transaccional, que almacena datos estructurados en grafos en lugar de en tablas" (Wikipedia, 2019).

IV-B. Algoritmos de detección de comunidades

Para llevar a cabo la detección de comunidades, se utilizaron distintos algoritmos de la librería "Graph Data Science" (GDS) de Neo4j a fin de poder comparar los resultados y proceder con el más convincente en términos de cantidad de grupos detectados y modularidad de la red, ya que los grafos con alta modularidad tienen conexiones densas entre los nodos dentro de las comunidades pero conexiones escasas entre los nodos en diferentes comunidades (Docs Neo4j Graph Data Science 2.4).

Algoritmo Louvain: Maximiza una puntuación de modularidad para cada comunidad, donde la modularidad cuantifica la calidad de una asignación de nodos a las comunidades. Esto significa evaluar qué tan densamente conectados están los nodos dentro de una comunidad, en comparación con qué tan conectados estarían en una red aleatoria (Blondel et al, 2008).

Algoritmo Label Propagation: Detecta estas comunidades utilizando únicamente la estructura de la red como guía y no requiere una función objetivo predefinida o información previa sobre las comunidades. LPA funciona propagando etiquetas a través de la red y formando comunidades basadas en este proceso de propagación de etiquetas. La intuición detrás del algoritmo es que una sola etiqueta puede convertirse rápidamente en dominante en un grupo de nodos densamente conectados, pero tendrá problemas para cruzar una región escasamente conectada (Docs Neo4j Graph Data Science 2.4).

Algoritmo Weakly Connected Components: Dos nodos están conectados, si existe un camino entre ellos. El conjunto de todos los nodos que están conectados entre sí forman un componente. A diferencia de los componentes fuertemente conectados (SCC), no se considera la dirección de las relaciones en la ruta entre dos nodos (Docs Neo4j Graph Data Science 2.4). Algoritmo Modularity Optimization: Intenta detectar comunidades en el grafo en función de su modularidad. Como se dijo anteriormente, la modularidad es una medida de la estructura de un grafo, midiendo la densidad de conexiones dentro de un módulo o comunidad. Los grafos con una puntuación de modularidad alta tendrán muchas conexiones dentro de una comunidad, pero solo unas pocas apuntarán hacia otras comunidades. El algoritmo explorará para cada nodo si su puntaje de modularidad podría aumentar si cambia su comunidad a uno de sus nodos vecinos (Docs Neo4j Graph Data Science 2.4).

IV-C. Test chi-cuadrado

Cuando se utiliza alguno de los algoritmos antes mencionados, es importante comprobar si las comunidades obtenidas están determinadas por características específicas observables en la base de datos. El test de chi cuadrado es una prueba estadística que se utiliza para determinar si existe una diferencia estadísticamente significativa entre la frecuencia esperada y las frecuencias observadas en una o más categorías de una tabla de contingencia. En las aplicaciones estándar de esta prueba, las observaciones se clasifican en clases mutuamente excluyentes.

La hipótesis nula en el test de chi cuadrado establece que no hay relación entre las variables categóricas y que cualquier diferencia observada se debe al azar. Si el valor obtenido del test de chi cuadrado es suficientemente grande, se rechaza la hipótesis nula y se concluye que existe una asociación significativa entre las variables. La ventaja de este test es que puede aplicarse a diferentes tipos de datos categóricos, como tablas de contingencia o frecuencias observadas en diferentes grupos.

IV-D. Modelos nulos

La aplicación de modelos nulos es una metodología ampliamente utilizada en estudios biológicos. Un modelo nulo consiste en una técnica estadística que genera patrones excluyendo deliberadamente ciertos mecanismos causales de interés (Gotelli, 2001), como puede ser por ejemplo la pertenencia a un grupo taxonómico. Esta aproximación estadística al análisis de patrones puede ser de una amplia gama de niveles de complejidad, pero en el presente trabajo se utilizó un modelo nulo relativamente simple con el objetivo de testear si las comunidades detectadas por el algoritmo Louvain están estructuradas con una señal taxonómica. Para esto se tomaron las comunidades con un mayor número de individuos siendo las primeras 8 para el caso de la red de hospedadores que comparten virus y las primeras 10 para la red de virus que comparten hospedadores. Luego para cada comunidad, se tomó al azar un número de nodos igual al número de nodos incluidos en la comunidad y se registraron las clases de estos nodos, obteniendo una distribución de frecuencias observadas para cada clase. Este proceso fue repetido 2000 veces para cada comunidad, obteniendo así la frecuencia media esperada y su intervalo de confianza del 95 % para cada una de las clases. Esto permite obtener una distribución nula de las clases en cada comunidad con las cuales comparar las observaciones. Si el número observado de una clase es mayor al esperado por el modelo nulo, se puede decir que esa comunidad tiene una mayor representación de dicha clase que lo esperado por azar, y por ende está siendo estructurada con una señal de su taxonomía. En el caso de la red de virus, esto reflejaría que los virus de la misma clase tienden a compartir huéspedes (recursos), existiendo así conservación de nicho a lo largo de la evolución.

V. Resultados

V-A. El grafo

Al cargar los distintos archivos que conforman la base de datos completa se encontraron 4.122 nodos tipo host, 9.753 nodos tipo virus y 25.403 vínculos "Infecta a" entre los mismos. En la Figura 3 se observa que hay un componente gigante totalmente conectado, un conjunto de 5 nodos y luego algunas tríadas y duplas de nodos aisladas.

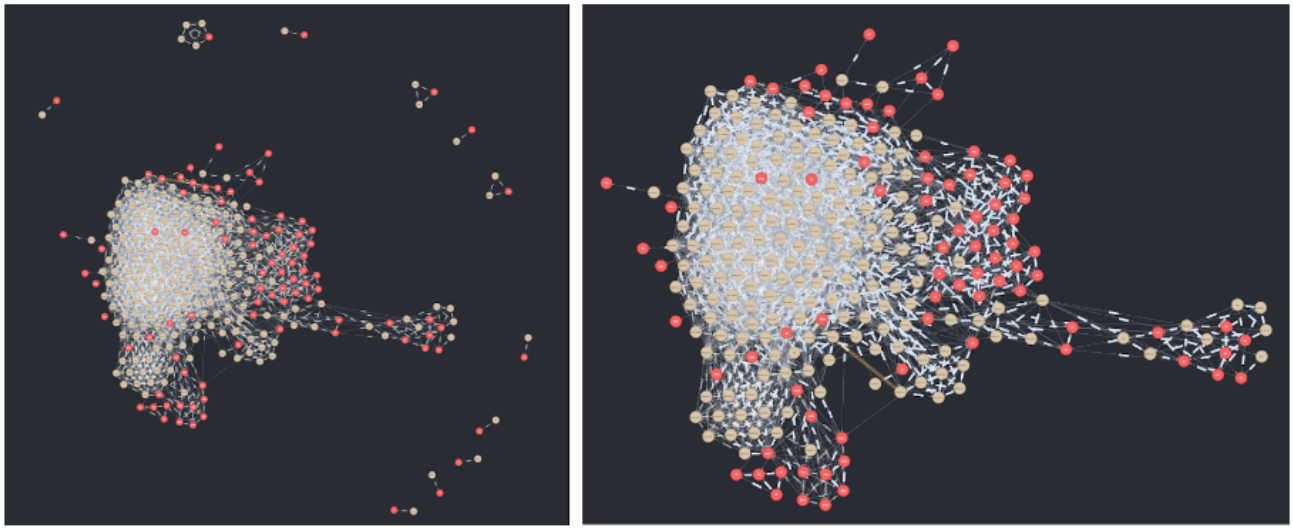

Figura 3. Red completa y red ampliada para la parte conexa (visualización con límite de 300 nodos, host=nodos rojos y virus=nodos beige)

	Total	Especies	Géneros	Familias	Órdenes	Clases
Host	4.122	3.673	1.774	479	116	9
Virus	9.753	9.518	370	67	36	27

Figura 4. Cantidad de nodos host y nodos virus por categoría

Al observar la Figura 5 y analizando la categoría de clasificación más amplia, es decir a nivel más macro (las clases), se puede afirmar que los mamíferos son los que priman dentro de los hospedadores (50 % del total de nodos) y dentro de los virus son los Pisoniviricetes (24 %), una clase de virus ARN monocatenario positivo. Yendo a un segundo nivel, se observan más nodos en el orden Rodentia que incluso está dentro de los mamíferos y más nodos Picornavirales que es un orden de virus que infectan animales, protistas y plantas

	Host			Virus		
	Órdenes	Nodos	%	Órdenes	Nodos	%
1	rodentia	517	14	picornavirales	1162	12
2	chiroptera	438	12	nidovirales	1114	12
3	passeriformes	393	11	mononegavirales	857	9
4	primates	235	6	herpesvirales	690	7
5	artiodactyla	232	6	NA	676	7
	Clases	Nodos	%	Clases	Nodos	%
1	mammalia	1841	50	pisoniviricetes	2276	24
2	aves	1207	33	monjiviricetes	861	9
3	actinopteri	660	18	herviviricetes	690	7
4	lepidosauria	165	4	NA	676	7
5	amphibia	162	4	ellioviricetes	624	7

Figura 5. Top 5 de órdenes y clases para cada tipo de nodo (% sobre el total de nodos de su tipo)

Tal como muestra la Figura 6, luego de los mamíferos, los hospedadores con mayor presencia en la fuente de datos son las aves. Esto podría dar un indicio de que los estudios de relaciones entre host y virus, y por ende la base de datos, está sesgada hacia los mamíferos, lo cual tiene sentido ya que es la clase a la que pertenecen los humanos y seguramente genere mayor interés a la hora de investigar. En el caso de los virus no se observa una tendencia tan marcada sino que la distribución es más pareja.

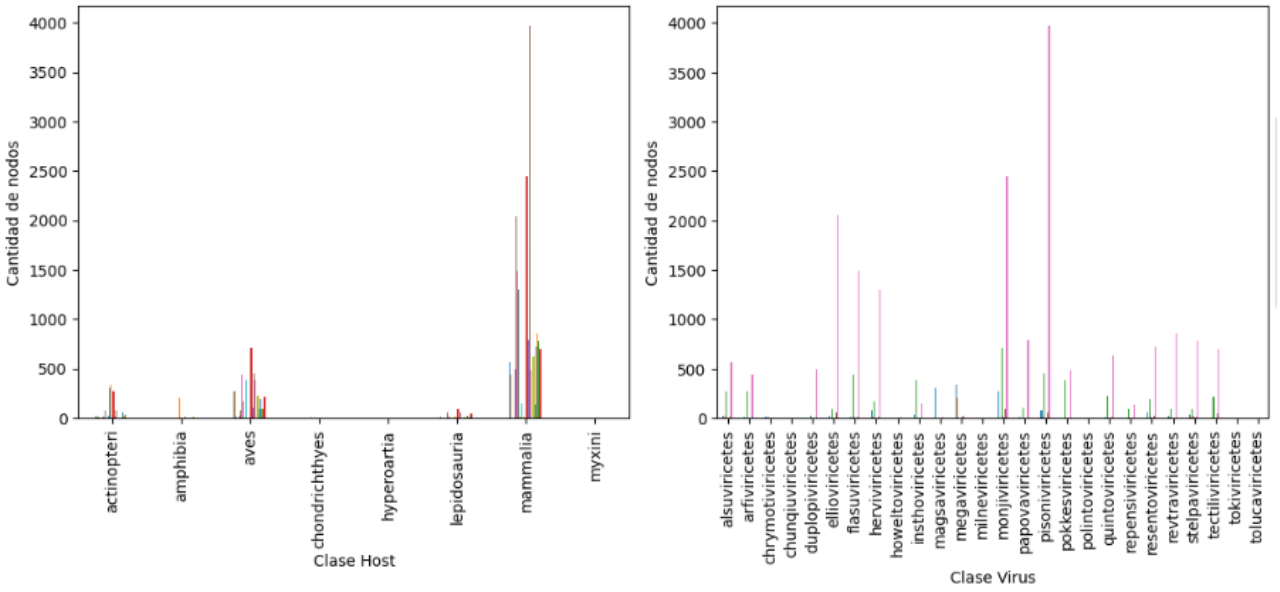

Figura 6. Cantidad de nodos de diferentes clases de virus para cada clase de host y viceversa

V-B. Riesgo zoonótico

A fin de investigar los virus que podrían representar una amenaza para la especie humana, se hace el siguiente estudio. Primero se seleccionan todos los virus que infectan a humanos según la base de datos VIRION. Luego se recuperan los

hospedadores que no son humanos pero que son infectados por esos virus. Finalmente, se seleccionan los otros virus que infectan a ese grupo de hospedadores relacionados con los humanos, es decir son virus que hoy en día no infectan a los humanos, por lo menos no directamente. A partir de estos datos se construye un ranking de virus que podrían infectar a los humanos debido a las conexiones antes mencionadas. La Figura 7 se lee de la siguiente manera: la primer especie de riesgo para los humanos es el virus bluetongue ya que infecta a 96 hospedadores que son infectados por algún virus que infecta al humano. El primer lugar está ocupado por una especie cuyo nombre no está disponible en la base pero que pertenece al género alphacoronavirus e infecta a 131 hospedadores que son infectados por algún virus que infecta al humano.

Especie	Género	Familia	Orden	Clase	Índice
NA	alphacoronavirus	coronaviridae	nidovirales	pisoniviricetes	131
bluetongue virus	orbivirus	reoviridae	reovirales	resentoviricetes	96
bat coronavirus	NA	coronaviridae	nidovirales	pisoniviricetes	85
carnivore protoparvovirus 1	protoparvovirus	parvoviridae	piccovirales	quintoviricetes	79
NA	orbivirus	reoviridae	reovirales	resentoviricetes	76

Figura 7. Top 5 especies de virus que podrían representar un riesgo para los seres humanos

A partir de la penúltima columna de la tabla, se concluye que las clases de virus pisoniviricetes y resentoviricetes podrían ser las más peligrosas para los humanos ya que en total infectan a 388 hospedadores que son infectados por algún virus que infecta al humano (si se analizan más del top 5 este número podría ser incluso mayor).

V-C. Medidas de centralidad

Existen diferentes medidas que describen la estructura topológica de la red, aquí el foco estará en las medidas de centralidad. Primero se halló la distribución de grado para cada tipo de nodo, encontrándose en ambos casos figuras que se asemejan a una distribución sesgada a la derecha o con asimetría positiva, ya que poseen una gran cantidad de nodos con grado 1 o 2, es decir 1 o 2 vínculos. Destaca la especie Homo Sapiens dentro de los hospedadores ya que es el único nodo con un grado comparativamente tan alto (1389), el resto de los nodos con mayor cantidad de vínculos rondan un grado igual a 500, tanto en los host como en los virus (ver Figura 8 y 9). La especie de virus con mayor cantidad de vínculos es Influenza A con 390. A su vez, considerando todos los nodos tipo host, el grado medio alcanzado es 5,9 y para los virus 2,3.

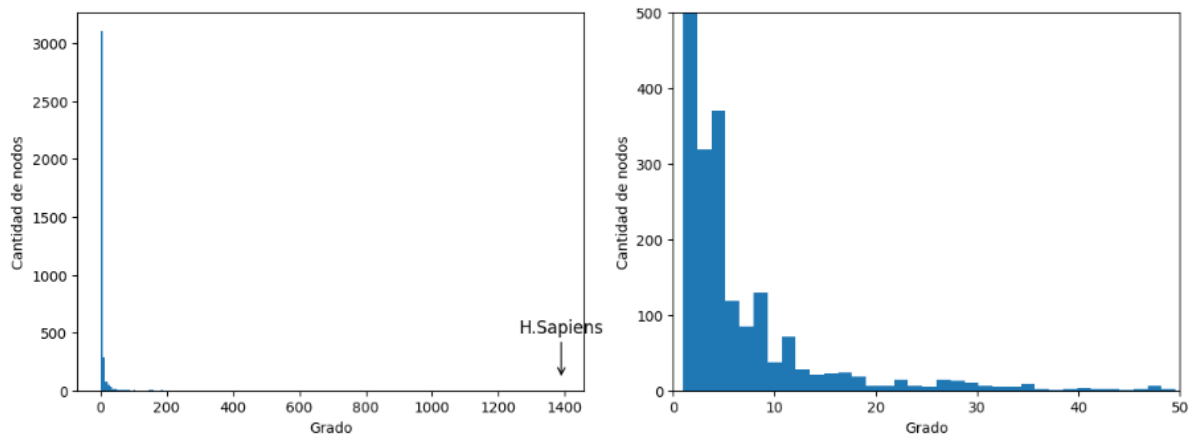

Figura 8. Distribución de grado de los hospedadores, versión ampliada a la derecha

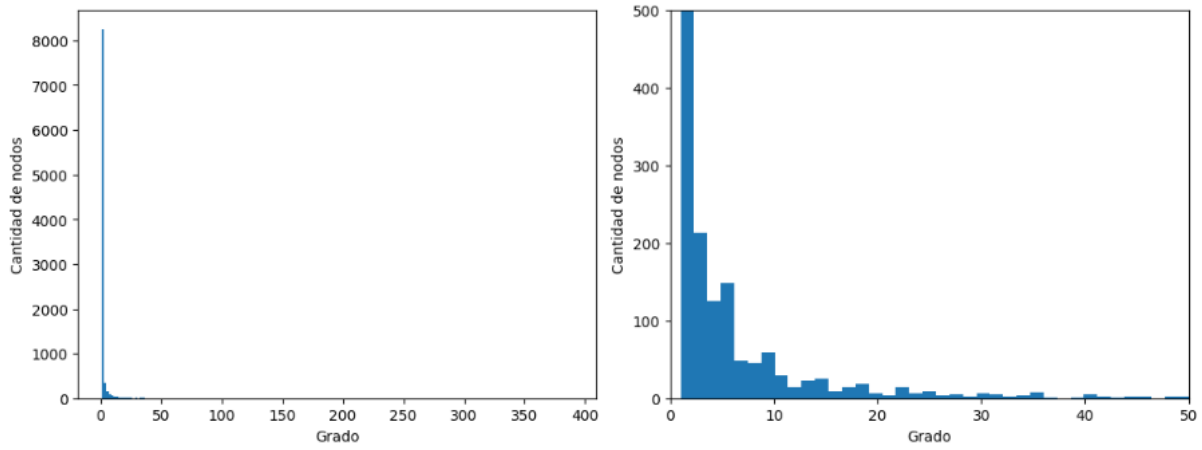

Figura 9. Distribución de grado de los virus, versión ampliada a la derecha

Estudiando la red en su totalidad, la densidad de la misma es 1,57. Esta medida refleja la proporción de vínculos existentes sobre el máximo de vínculos posibles.

Además, se calcularon algunas medidas de centralidad presentes en la librería GDS antes mencionada. El algoritmo Page Rank mide la importancia de cada nodo dentro del grafo, en función del número de relaciones entrantes y la importancia de los nodos de origen correspondientes. La suposición subyacente, en términos generales, es que una página es tan importante como las páginas que la vinculan. Por otro lado, el Article Rank es una variante del algoritmo Page Rank, que mide la influencia transitiva de los nodos. Page Rank sigue la suposición de que las relaciones que se originan en los nodos de bajo grado tienen una mayor influencia que las relaciones de los nodos de alto grado. Article Rank reduce la influencia de los nodos de bajo grado al reducir las puntuaciones que se envían a sus vecinos en cada iteración.

Una vez implementados ambos algoritmos, los resultados obtenidos fueron exactamente los mismos en cuanto al orden de los nodos, lo cual implica que las relaciones que se originan en los nodos de bajo grado en este caso no tienen una mayor influencia que las relaciones de los nodos de alto grado. A continuación se muestra la centralidad de Page Rank donde como era esperable los nodos de cada tipo con mayor cantidad de vínculos obtuvieron el primer lugar.

Nodos Host		Nodos Virus	
Especie	Centralidad	Especie	Centralidad
homo sapiens	395,38	influenza a virus	98,73
macaca mulatta	188,35	avian orthoavulavirus 1	63,57
sus scrofa	103,23	west nile virus	54,04

Figura 10. Resultados centralidad Page Rank

VI. Detección de comunidades

Previo a la implementación de los diferentes algoritmos para detectar comunidades, se debió agregar la red al catálogo de grafos de la librería GDS, para ello se creó la RED1 cuya característica principal es que no es dirigida y por lo tanto los 25.403 vínculos “Infecta a” presentes en el grafo dejan de tener una dirección; y la RED2 en la cual se determinó que los vínculos no sólo tienen dirección sino que se producen en ambos sentidos, es decir que además del link “Infecta a” están los mismos en el sentido contrario con el nombre “Infectado por” (25.403*2=50.806 vínculos en total). A partir de la ejecución del algoritmo Louvain para ambas redes, se determinó seguir trabajando con la RED1 ya que la modularidad obtenida fue mucho menor en la RED2 (0,74 vs 0,49 respectivamente).

Tal como se puede ver en la Figura 11, el método Modularity Optimization determina una cantidad de comunidades notoriamente mayor en comparación al resto pero dado que se va a trabajar a nivel de clase dentro de las estructuras taxonómicas y existen 9 para los nodos tipo Host y 27 para los nodos tipo Virus, se decide no considerar los resultados de este método ni

los de Label propagation y analizar los otros. Los algoritmos Louvain y Weakly Connected Components detectan una cantidad similar de grupos pero siguiendo el criterio de obtener la mayor modularidad posible se decide avanzar con el primer algoritmo.

Algoritmo	Cant. comunidades	Modularidad[6]
Louvain	439	0,74
Label Propagation	771	-
Weakly Connected Components	409	-
Modularity Optimization	1742	0,56

Figura 11. Detección de comunidades con diferentes algoritmos

Analizando las clases predominantes en las 5 comunidades con mayor cantidad de nodos, se puede afirmar que desde la perspectiva de los virus, a excepción de una comunidad, en el resto se observa una mayor presencia de la clase pisoniviricetes. En el caso de los hospedadores, sucede algo parecido, en uno de los grupos priman las aves pero en el resto la clase mayoritaria dentro de los host es la de los mamíferos. Al mismo tiempo, en todos los grupos se observa que el peso de los nodos de la clase mayoritaria no supera el tercio de los nodos totales en la comunidad de pertenencia (ver Figura 12).

Comunidad	1	2	3	4	5
ID Comunidad	13445	4687	4167	8281	4300
Nodos totales	1906	1606	990	936	926
Clase mayoritaria H.	mammalia	aves	mammalia	mammalia	mammalia
Clase mayoritaria V.	pisoniviricetes	pisoniviricetes	pisoniviricetes	ellioviricetes	pisoniviricetes
% clase mayorit. H. en el total	14%	36%	1%	12%	12%
% clase mayorit. V. en el total	35%	8%	25%	8%	7%

Figura 12. Nodos totales y clases de Host y Virus para las 5 comunidades más densas

VI-A. Test Chi-Cuadrado

Para las 5 comunidades con más nodos, el test de χ^2 indica que se rechaza la hipótesis nula (ver Figura 13), por lo que se puede afirmar que existe una diferencia estadísticamente significativa entre la distribución de frecuencias observadas y la esperada por azar en las 5 comunidades.

	Comunidad 1	Comunidad 2	Comunidad 3	Comunidad 4	Comunidad 5
p-valor	3,439e-144	9,072e-170	2,011e-105	5,790e-38	2,034e-33

Figura 13. Resultados test chi-cuadrados par las primeras 5 comunidades con más nodos

[6]No se logró calcular la modularidad para los algoritmos Label Propagation y Weakly Connected Components.

VI-B. Proyección de la red bipartita

La idea al proyectar la red de Host y de Virus era poder entender mejor las relaciones y asociaciones entre los nodos de cada tipo. Es por eso que se hizo la proyección de cada una (ver Figura 14) y luego se le aplicó el mismo algoritmo de detección de comunidades explicado anteriormente, Louvain (ver Figura 15).

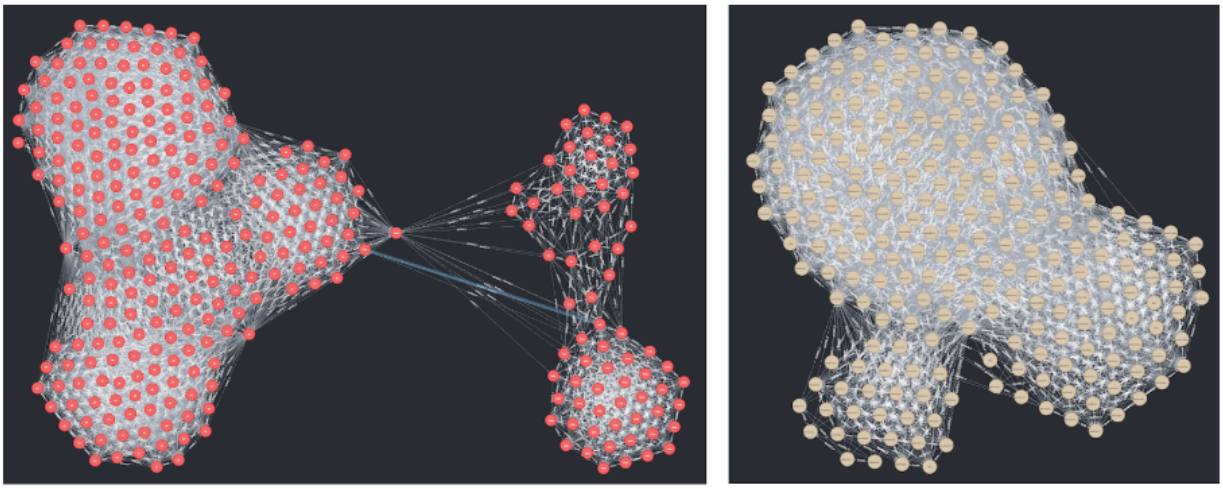

Figura 14. Proyecciones redes Host y Virus (visualizaciones con límite de 300 nodos c/u)

En comparación a los resultados obtenidos anteriormente, el mayor cambio se observa en las comunidades en la red de hospedadores ya que aparecen grupos diferenciados por clase de hospedador, dos mayoritariamente de mamíferos, dos en los que priman las aves y uno de peces óseos o lo que es lo mismo peces dotados de esqueleto interno (actinopteri). Este resultado también está alineado con las tres clases de mayores proporciones en la base de datos, son las que tienen mayor cantidad de animales.

Comunidades Host	1	2	3	4	5
ID Comunidad	1294	36	97	516	349
Nodos totales	1405	690	654	264	254
Clase mayoritaria	mammalia	aves	actinopteri	aves	mammalia
% clase en total	91%	90%	73%	66%	94%
Comunidades Virus	1	2	3	4	5
ID Comunidad	906	29	45	653	79
Nodos totales	2239	1450	1256	1080	1027
Clase mayoritaria	pisoniviricetes	pisoniviricetes	pisoniviricetes	pisoniviricetes	ellioviricetes
% clase en total	41%	16%	40%	15%	23%

Figura 15. Nodos totales y clases de Host y Virus para las 5 comunidades más densas dentro de cada red proyectada

VI-C. Modelos nulos

Para todas las comunidades de hospedadores analizadas, se encontró que todas las clases están fuera de los intervalos de confianza arrojados por los modelos nulos, o sea, con un número de nodos por clase con mayor o menor representación que lo esperado si las comunidades estuvieran ensambladas al azar. Para las 8 comunidades analizadas existían entre 1 y 2 clases con números mayores a los esperados por el modelo nulo y el resto de las clases presentaban valores menores a los esperados por el modelo nulo. Sin embargo, el modelo nulo estimaba valores cercanos a 0 (aunque estrictamente mayores) para la clase hyperoartia en ciertas comunidades, donde su valor observado fue 0, por lo que por la naturaleza de los datos (conteos) podría considerarse que estaba dentro de la estimación. Las hyperoartias (conocidas vulgarmente como lampreas)

estaban poco representadas en la base de datos. Es posible que esta baja representación en la base de datos explique por qué su representación en las comunidades detectadas sea en números similares a los esperados por el modelo nulo. Si la base de datos contiene poca información de la estructura de las relaciones entre los virus y las lampreas es posible que no sea representativa de la clase en general, por lo que las pocas lampreas en la base de datos podrían terminar dispersas en distintas comunidades en lugar de estar agrupadas.

En el caso de los virus los resultados son ligeramente diferentes. Para todas las comunidades analizadas se observaron sobre representaciones respecto al modelo nulo así como subrepresentaciones, pero la cantidad de clases con sobre representaciones es mayor (6 en promedio), y aparecen varias comunidades con una representación de ciertas clases dentro de lo esperado por el modelo nulo. Esto refleja una menor estructura taxonómica en las comunidades detectadas en la red de virus. Es posible que la naturaleza de la recolección de estos datos genere este tipo de patrones, ya que por lo general se estudian colecciones de individuos de vertebrados de cierto grupo taxonómico y se estudian sus virus, mientras que es menos frecuente que el proceso de recolección de la información se de en el sentido inverso.

Clase	Media Esperada	Límite inferior	Límite superior	Observados
mammalia	491,99	480,55	503,44	1270,0
aves	291,93	285,32	298,54	90,0
actinopteri	138,08	134,92	141,24	5,0
lepidosauria	30,12	29,39	30,85	7,0
amphibia	38,83	37,73	39,93	1,0
NA	11,42	11,11	11,73	5,0
chondrichthyes	2,97	2,88	3,05	0,0
hyperoartia	0,67	0,64	0,70	0,0

Figura 16. Ejemplo del resultado del modelo nulo para una comunidad de la red proyectada de hospedadores. Pueden verse las medias esperadas para cada clase de hospedadores y el intervalo de confianza de las mismas, y en la última columna el número de nodos observados en la comunidad de cada clase. En este caso puede observarse que esta comunidad presenta más mamíferos que lo esperado por el modelo nulo, mientras que el resto de las clases está subrepresentado en relación a los resultados esperados por modelo nulo.

VII. Conclusiones y Reflexiones Finales

Tras llevar a cabo el presente trabajo, se puede afirmar que se ha logrado migrar la base VIRION a una base de datos en grafo que siga el Property Graph Model así como realizar consultas a la misma de modo de obtener información sobre las características de los nodos y su distribución. Al mismo tiempo, se pudieron aplicar diferentes técnicas y/o metodologías para explorar la base, entre ellas la detección de comunidades y los modelos nulos.

A partir de los algoritmos de detección de comunidades aplicados en la red general y en la red proyectada para los virus, se encontró que los grupos a priori no están diferenciados por clases de nodos. Sin embargo, en la red proyectada de hospedadores aparecen grupos separados por clase de hospedador, dos mayoritariamente de mamíferos, dos en los que priman las aves y uno de peces óseos o lo que es lo mismo peces dotados de esqueleto interno (actinopterygii). Estos resultados preliminares están alineados con el peso de cada clase en el total de nodos de la base.

Los resultados obtenidos tanto en los modelos nulos y los test estadísticos indican que, al menos para el máximo nivel taxonómico encontrado en la base de datos, efectivamente los grupos de virus u hospedadores que tienden a estar más vinculados entre sí representan ciertos grupos taxonómicos, por lo que podemos decir que existe una tendencia a la conservación de nicho a lo largo de la evolución en esta característica. En este sentido el presente trabajo permite discernir entre las hipótesis contrastantes que presenta el marco teórico, pero solo a la escala de resolución taxonómica de las clases. Para profundizar en el contraste de hipótesis sería necesario realizar análisis similares a otras resoluciones taxonómicas, e incluir mayores restricciones en el diseño de los modelos nulos como pueden ser el factor biogeográfico de los distintos grupos taxonómicos, o la maquinaria bioquímica viral.

En cuanto al riesgo zoonótico, se puede afirmar que la primer especie de riesgo para los humanos es el virus bluetongue ya que infecta a 96 hospedadores que son infectados por algún virus que infecta al humano. Además se concluye que las

clases de virus pisoniviricetes y resentoviricetes podrían ser las más peligrosas para los humanos ya que en total infectan a 388 hospedadores que son infectados por algún virus que infecta al humano (si se analizan más del top 5 este número podría ser incluso mayor).

Como reflexión final vale destacar que los métodos empleados fueron satisfactorios y acordes al problema planteado. Sin embargo, la coexistencia en la base de datos de dos estructuras, una donde las relaciones son entre virus y hospedador, y otra donde los nodos de una misma etiqueta están relacionados entre sí en función de si comparten virus/hospedador (redes proyectadas), es probablemente ineficiente, ya que la información es redundante, pero permite hacer consultas más fáciles de ejecutar.

Referencias

[1] Vincent D Blondel et al. "Fast unfolding of communities in large networks". En: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (9 de oct. de 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. URL: https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008 (visitado 09-07-2023).

[2] Colin J. Carlson et al. "The Global Virome in One Network (VIRION): an Atlas of Vertebrate-Virus Associations". En: *mBio* 13.2 (26 de abr. de 2022). Ed. por Brett E. Pickett y Kellie Jurado, e02985-21. ISSN: 2150-7511. DOI: 10.1128/mbio.02985-21. URL: https://journals.asm.org/doi/10.1128/mbio.02985-21 (visitado 09-07-2023).

[3] *Community detection - Neo4j Graph Data Science*. Neo4j Graph Data Platform. URL: https://neo4j.com/docs/graph-data-science/2.4/algorithms/community/ (visitado 09-07-2023).

[4] Nicholas J. Gotelli. "Research frontiers in null model analysis: *Null model analysis*". En: *Global Ecology and Biogeography* 10.4 (jul. de 2001), págs. 337-343. ISSN: 1466822X. DOI: 10.1046/j.1466-822X.2001.00249.x. URL: http://doi.wiley.com/10.1046/j.1466-822X.2001.00249.x (visitado 09-07-2023).

[5] *Graph Data Science*. Graph Database & Analytics. URL: https://neo4j.com/product/graph-data-science/ (visitado 09-07-2023).

[6] *Modularity metric - Neo4j Graph Data Science*. Neo4j Graph Data Platform. URL: https://neo4j.com/docs/graph-data-science/2.4/algorithms/alpha/modularity/ (visitado 09-07-2023).

[7] *Neo4j*. En: *Wikipedia, la enciclopedia libre*. Page Version ID: 118034594. 6 de ago. de 2019. URL: https://es.wikipedia.org/w/index.php?title=Neo4j&oldid=118034594 (visitado 09-07-2023).