Journal of
Information Systems Engineering
and Business Intelligence

Vol.10, No.1, February 2024
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Leveraging Biotic Interaction Knowledge Graph and Network Analysis to Uncover Insect Vectors of Plant Virus

Moh. Zulkifli Katili [1)*], Yeni Herdiyeni [2)*], Medria Kusuma Dewi Hardhienata [3)]

[1)2)3)] *Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia*

[1)]moh.zulkiflikatili@apps.ipb.ac.id, [2)]yeni.herdiyeni@apps.ipb.ac.id, [3)]medria.hardhienata@apps.ipb.ac.id

Abstract

Background: Insect vectors spread 80% of plant viruses, causing major agricultural production losses. Direct insect vector identification is difficult due to a wide range of hosts, limited detection methods, and high PCR costs and expertise. Currently, a biodiversity database named Global Biotic Interaction (GloBI) provides an opportunity to identify virus vectors using its data.
Objective: This study aims to build an insect vector search engine that can construct an virus-insect-plant interaction knowledge graph, identify insect vectors using network analysis, and extend knowledge about identified insect vectors.
Methods: We leverage GloBI data to construct a graph that shows the complex relationships between insects, viruses, and plants. We identify insect vectors using interaction analysis and taxonomy analysis, then combine them into a final score. In interaction analysis, we propose Targeted Node Centric-Degree Centrality (TNC-DC) which finds insects with many directly and indirectly connections to the virus. Finally, we integrate Wikidata, DBPedia, and NCBIOntology to provide comprehensive information about insect vectors in the knowledge extension stage.
Results: The interaction graph for each test virus was created. At the test stage, interaction and taxonomic analysis achieved 0.80 precision. TNC-DC succeeded in overcoming the failure of the original degree centrality which always got bees in the prediction results. During knowledge extension stage, we succeeded in finding the natural enemy of the *Bemisia Tabaci* (an insect vector of *Pepper Yellow Leaf Curl Virus*). Furthermore, an insect vector search engine is developed. The search engine provides network analysis insights, insect vector common names, photos, descriptions, natural enemies, other species, and relevant publications about the predicted insect vector.
Conclusion: An insect vector search engine correctly identified virus vectors using GloBI data, TNC-DC, and entity embedding. Average precision was 0.80 in precision tests. There is a note that some insects are best in the first-to-fifth order.

Keywords: Knowledge Graph, Network Analysis, Degree Centrality, Entity Embedding, Insect Vector

Article history: Received 22 September 2023, first decision 9 January 2024, accepted 5 February 2024, available online 28 February 2024

I. Introduction

In plant ecosystems, insects have role as vectors that carry viruses. Several insect vectors [1–3] spread viruses in plants and caused significant losses in agricultural production [4–7]. Since insects are responsible for the majority of virus transmission [2], the role of insect vectors is crucial. Therefore, identifying insect vectors is important to break the transmission and reduce its negative impact on agricultural production.

Collecting insects near virus-infected plants and identifying them in a lab are typical methods for insect vector identification [8, 9]. However, this method poses several challenges, such as the ability of some insects to infect multiple hosts (wide host range) [10], limited detection methods [11], and a process that requires expertise in implementation [11, 12]. Literature screening of prior studies [1–3] is another method for identifying insect vectors, given the abundance of prior studies [1–3] that have published information relating to insect vectors. However, reading the literature systematically and summarizing all the information as a whole also requires more time and effort. On the other hand, agricultural knowledge management methods have developed knowledge graph ontology technology [13]. Knowledge graphs have two advantages: First, because of the Linked Open Data (LOD) principle [14], which makes the data connected with other sources of knowledge; and second, due to the graph format, which enables the application of network analysis techniques [15]. One of the large knowledge graphs that can be used to identify insect vectors is Global Biotic Interaction (GloBI) [16]. GloBI is a global database that stores interaction data for all entities or living things in their ecosystem, including the interactions of insects, viruses, and plants.

Previous research has used ontologies to address plant pathology problems [17–21]. However, research that applies network analysis to GloBI graphs to identify insect vectors has not been widely carried out. As suggested by [22],

[*] Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print) © 2024 The Authors. Published by Universitas Airlangga.

doi: http://dx.doi.org/10.20473/jisebi.10.1.94-109

applying network analysis to identify insect vectors is a potential practice. Our study aims to build a vector insect search engine based on network analysis. The search engine must be able to: (1) construct a knowledge graph of virus-insect-plant interactions; (2) identify insect vectors using network analysis; and (3) collect detailed information regarding identified insect vectors.

To construct graphs, we use the Breadth-First Search (BFS) technique [23] to retrieve as much interaction data as possible from the GloBI web service. Then the data is preprocessed and converted into the graph format. To identify insect vectors, we use two analytical approaches: interaction analysis and taxonomic analysis. Interaction analysis is based on Lilies' insect vector identification method [8], which involves examining insects that interact with virus-infected plants in the field. Meanwhile, taxonomic analysis is based on research findings [1, 3], which show that certain insect families have unique association patterns with various virus genera. In interaction analysis, we proposed the Targeted Node Centric-Degree Centrality (TNC-DC). TNC-DC is a modification of the Degree Centrality technique [24] that has been adapted to search for central nodes while considering the relationship with a target node (in this case, a virus). In taxonomic analysis, we use entity embedding techniques [25] to convert taxonomic data and Euclidean distance [26] to find out the similarity taxa of each node. Then the two approaches' results are combined to produce a single final score. The higher the final score, the greater the potential of the insect as a vector. To extend our knowledge about predicted insects, we use some knowledge bases named DBpedia [27], Wikidata [28], and NCBI Taxon Ontology [29, 30] to obtain additional pertinent information. Included in this information are the insect's common name, specifics about its taxonomy, and an explanation of how it interacts with its natural biocontrol.

Ultimately, the three procedures explained before were used as fundamental stages in the insect vector search engine that we have built, namely Vektorpedia. Researchers and other stakeholders can use Vektorpedia to understand the dynamics of virus-insect-plant interactions and formulate insect pest control strategies.

II. Literature Review

According to Hogan [31], there are two general techniques for knowledge accumulation in knowledge graphs: deductive (rule and logical inference) and inductive (machine learning and network analysis). In the field of plant pathology, most previous studies have used logical inference and ontology coverage on certain plants, such as cocoa plants [17], rice [18], and several other plants [19–21]. While research using inductive techniques and covering a wide range of organisms is relatively rare. The study [32] used a wide range of organismal data from GloBI to track parasites. However, it did not mention the identification of insect vectors. The study [33] specifically addresses the interaction of insects and host plants using network analysis but aims to identify association patterns between insects and their hosts.

To the best of our knowledge, no research has been found that applies network analysis to GloBI graphs to assist in identifying insect vectors. On the other hand, applying network analysis to help identify insect vectors is a potential practice because it has been suggested in researches [22, 34]. According to Garret [22], plant pathology problems are complex system problems, and network analysis [15] can overcome these problems. Several previous studies have suggested applying network analysis to plant pathogen problems, especially using information from species interaction data to identify viral vectors [22, 34, 35]. We implement one of Garrett's suggestions, solving the problem of insect vectors with network analysis. We utilize the virus-insect-plant interaction data already in GloBI [16] and apply network analysis to find insect vectors. A comparison of several studies related to the research to be submitted can be seen in Table 1.

TABLE 1
Comparison Between Our Research and Previous Research

Work	Approach	Objective
Lagos-Ortiz et al, 2019	Deductive: *logical inference*	Identification of plant diseases and pests
Jearanaiwongkul et al, 2021	Deductive: *logical inference*	Identification of plant diseases
Rodríguez-García et al, 2021	Deductive: *logical inference*	Identification of plant diseases and pests
Seltmann et al. 2020	-	Tracking parasites on a terrestrial basis
Our research	Inductive: *targeted node centric-degree centrality* and *entity embedding*	Identification of plant virus vectors

III. Methods

The research stages consist of: 1) Data acquisition 2) graph conversion; 3) interaction analysis; 4) taxonomic analysis; 5) final-score calculation; 6) testing; 7) knowledge extension; and 8) application development. An overview of the research stages is shown in Fig. 1. Knowledge graphs and an ontology are used as data sources in data acquisition, knowledge extension, and taxonomic analysis. The main data is GloBI, and the complementary data are NCBI Taxon Ontology, Wikidata, and DBPedia. Detailed specifications and data usability can be seen in Table 2.

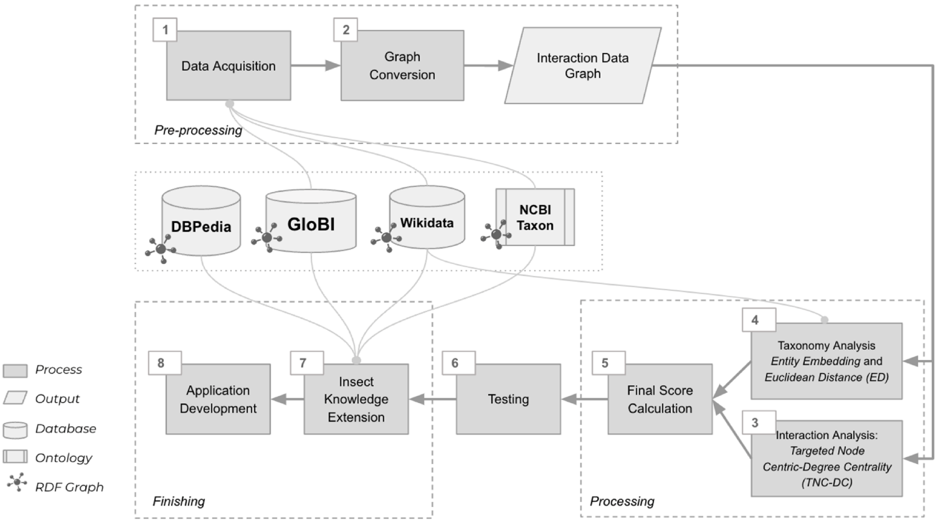


Fig. 1 Research stages

TABLE 2
RESEARCH DATA

Title	Format	Purpose
GloBI	Linked Open Data	As the main data of biotic interactions
NCBITO	Ontology	As complementary data on the taxonomy of plants, insects, and viruses
Wikidata	Linked Open Data	As a hub for disambiguation of taxonomic data, and detailed insect detail data
DBPedia	Linked Open Data	As complementary details of insect data

A. Data Acquisition

Data acquisition was performed twice: once for virus interactions (Depth 1) and then again for some insects or plants that are related to the virus (Depth 2). The process of acquiring data is broken down into the following sub-stages: (1) data retrieval; (2) table separation; (3) disambiguation; (4) table concatenation; and (5) Filter are the five steps in this process. Fig. 2 shows all the sub-stages.

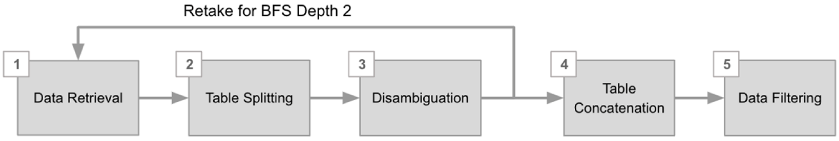


Fig. 2 Sub-stages in data acquisition

1) Data Retrieval

This stage aims to expand and collect all possible virus, insect, and plant interactions in GloBI. The Breadth-First Search (BFS) technique [23] is utilized for retrieving interaction data. BFS was chosen due to its capability to retrieve all organisms associated with the virus. The concept of data retrieval using BFS is illustrated in Fig. 3.

Since the goal is to identify the virus's insect vector, BFS is initiated with the virus itself as the starting node. The BFS is conducted at two depths. In Depth 1 (first data acquisition), the objective is to acquire plants and insects directly linked to the virus. In Depth 2 (second data acquisition), the aim is to obtain other entities connected to viruses indirectly through plants or insects obtained from Depth 1. Depth 2 consists of plant BFS as the host (Depth 2.1), insect BFS as the pathogen (Depth 2.2), and insect BFS as the host (Depth 2.3). In the end, this stage generates two raw tabular tables. One from Depth 1 and the other from Depth 2.

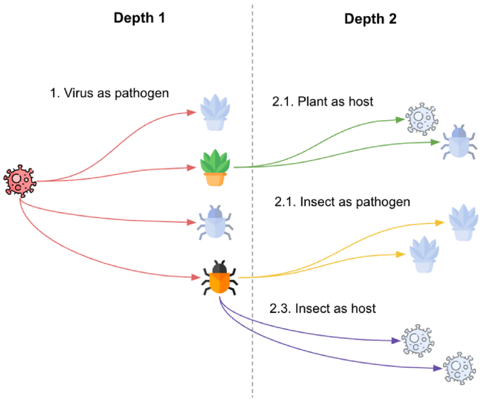
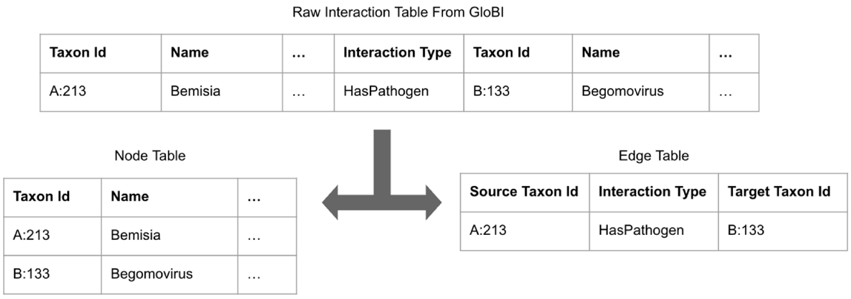


Fig. 3 The concept of virus-insect-plant interaction separated into two depths

2) *Table Splitting*

This stage aims to separate the raw interaction table from GloBI into a node table and an edge table. GloBI raw data, as shown in Fig. 4, is a table consisting of interaction records. Each row contains information on two entities and the type of interaction. For this reason, data needs to be separated between tables containing entity information (nodes) and tables containing interaction information (edges). This separation aims to eliminate redundancies and data duplication. Fig. 4 shows the table splitting process.


Raw Interaction Table From GloBI

Taxon Id	Name	...	Interaction Type	Taxon Id	Name	...
A:213	Bemisia	...	HasPathogen	B:133	Begomovirus	...

Node Table

Taxon Id	Name	...
A:213	Bemisia	...
B:133	Begomovirus	...

Edge Table

Source Taxon Id	Interaction Type	Target Taxon Id
A:213	HasPathogen	B:133

Fig. 4 Table Splitting

3) *Disambiguation*

At this stage, ambiguity in an entity (an individual virus, insect, or plant) is eliminated to determines the most appropriate taxon name for an entity based on the NCBI Taxonomy Ontology (NCBITO). The process includes matching an entity with a standard entity from the NCBI Taxonomy Ontology via a big knowledge base as a hub (Wikidata). Fig. 5 shows the process of this stage.

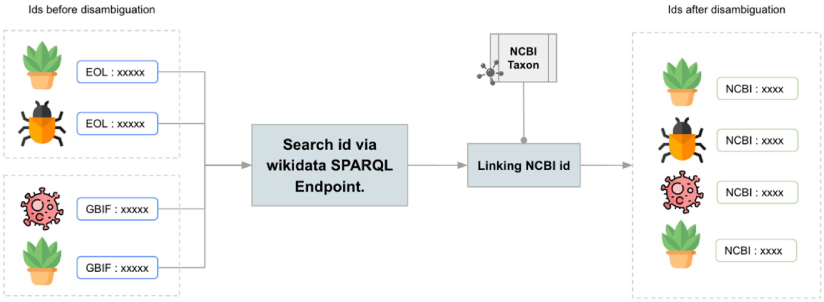
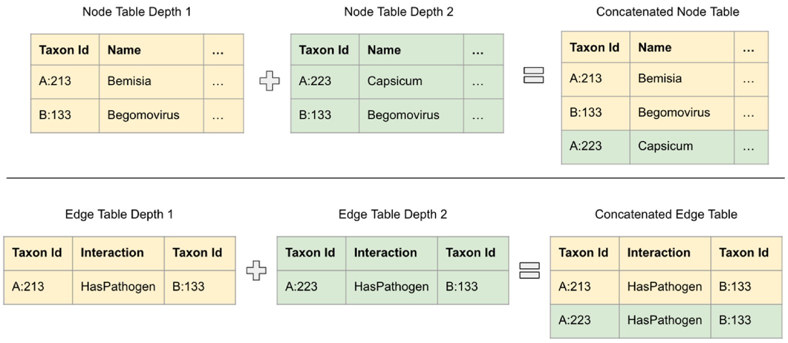


Fig. 5 Disambiguation process to standardize taxon IDs

In the first step, entities are grouped based on their biodiversity database prefix code, excluding the NCBITO database code. Then, each entity within each group is linked to entities in Wikidata by matching their IDs. Once the linked entities from Wikidata are obtained, possibly NCBI IDs associated with them are retrieved. This matching process utilizes an ontology query language endpoint known as "Simple Protocol and RDF Query Language" (SPARQL) provided by Wikidata. After obtaining NCBI IDs, they are linked again to the NCBI Taxon Ontology in order to obtain the entire taxon path. This entire process is carried out iteratively for each ID group based on the same database code. Finally, all retrieved data is rechecked, and any duplicates with existing data are deleted.

4) Table Concatenation

As depicted in Fig. 2, this stage is carried out after the data retrieval, table splitting, and disambiguation for the Depth 1 and Depth 2 interactions have been completed. The objective of this stage is to concatenate the interaction data acquired from the Depth 1 and Depth 2 processes. In this stage, both the node and edge tables from BFS Depth 1 and Depth 2 are concatenated. The table-concatenation process is illustrated in Fig. 6.



Node Table Depth 1

Taxon Id	Name	...
A:213	Bemisia	...
B:133	Begomovirus	...

Node Table Depth 2

Taxon Id	Name	...
A:223	Capsicum	...
B:133	Begomovirus	...

Concatenated Node Table

Taxon Id	Name	...
A:213	Bemisia	...
B:133	Begomovirus	...
A:223	Capsicum	...

Edge Table Depth 1

Taxon Id	Interaction	Taxon Id
A:213	HasPathogen	B:133

Edge Table Depth 2

Taxon Id	Interaction	Taxon Id
A:223	HasPathogen	B:133

Concatenated Edge Table

Taxon Id	Interaction	Taxon Id
A:213	HasPathogen	B:133
A:223	HasPathogen	B:133

Fig. 6 Node and edge tables are vertically concatenated

Each table (both node and edge) is concatenated vertically based on rows of data. This concatenation facilitates the identification of duplicated data from both Depth 1 and Depth 2. If there is data with the same ID, the duplication can be readily identified and subsequently deleted.

5) Filtering

This stage is the final data cleaning before starting to convert the graph. It consists of removing duplicates and filtering the data to exclusively capture viruses, insects, and plants. Other entities except the virus, insect, and plant are dropped. This leads to a slight decrease in data. However, it is essential for ensuring data consistency and enhancing the accuracy of subsequent analyses.

B. Graph Conversion

Since network analysis can only be applied to graph data, the data must be converted into graph format. The graph conversion process is using NetworkX [36]. NetworkX is a Python package containing functions used for graph conversion. It takes a node table and an edge table as input and produces a directed tripartite graph as output. This tripartite graph is constructed with nodes representing viruses, insects, and plants. The edges connecting these nodes are described by relationship types such as hasHost, pathogenOf, pollinates, visitFlowersOf, or visit. Fig. 7 illustrates the resulting tripartite graph created during this process.

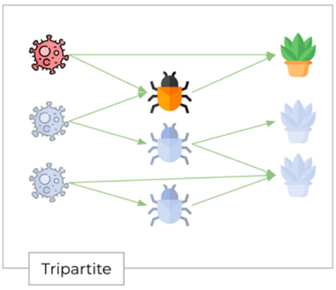


Fig. 7 Tripartite graph with three class (viruses, insects and plants)

C. Interaction Analysis

Interaction analysis uses the Targeted Node Centric-Degree Centrality (TNC-DC) technique. TNC-DC is a modification of Freeman's degree centrality [24] that we proposed. The difference between TNC-DC and Freeman's degree centrality is that there is a multiplier weight based on the number of direct relationships between the calculated node (insect) and the target node (virus), or at least there are indirect relationships through other entities first (e.g., plants). Fig. 8 shows how the idea of TNC-DC was used in the analysis of interactions.

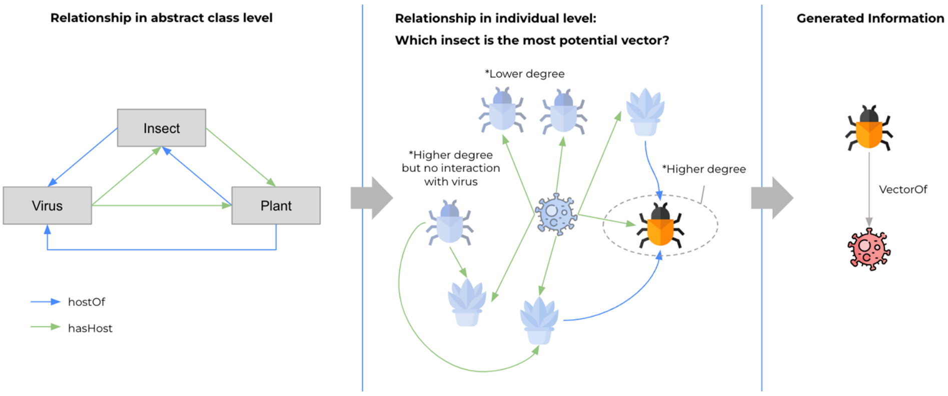

Relationship in abstract class level
Insect
Virus
Plant
hostOf
hasHost
Relationship in individual level:
Which insect is the most potential vector?
*Lower degree
*Higher degree but no interaction with virus
*Higher degree
Generated Information
VectorOf

Fig. 8 The degree centrality concept to be applied in interaction analysis

In general, the data have a relationship concept at an abstract level consisting of three entity classes (namely viruses, insects, and plants) and two relationship types (namely hasPathogen and pathogenOf). At the individual level, the graph is more complex. However, the primary objective is to identify insects that fulfill specific criteria: (1) have the most relationships (highest degree), (2) have direct relationships with the most viruses, and (3) have indirect relationships with viruses the most. In the end, insects that meet these requirements have a high degree of relationship and are assumed to be the insects with the most potential as viruses. A list of potential vector insects would be the information produced at this stage. The TNC-DC formula can be seen in equation 1.

$$C_{TNCDC}(p_k) = \left(\frac{\sum_{i=1}^{n} a(p_i, p_k)}{n-1} \times \frac{\sum_{i=1}^{n} w(p_i, p_k)}{n-1} \times \frac{1+\sum_{i=1}^{n} o(p_i, p_k)}{(n-2)} \right) \quad (1)$$

The formula consists of three parts, namely the original degree calculation (a), the direct virus relationship calculation (w), and the indirect virus relationship calculation (o). The p_k is the node to be calculated, p_i is the other connected nodes, and n is the total node. $a(p_i,\ p_k) = 1$ only if there are connected edges. $w(p_i,\ p_k) = 1$ only if there is an edge that is directly connected to the node and p_i is the main virus node. $o(p_i,\ p_k) = 1$ only if there are connected edges, but through another node first. In the o calculation, the constant 1 is added so that if the value of $o = 0$, then the values of a and w do not become 0 either. The addition of 1 then has implications for the total divisor, which must be changed to $(n - 2)$ to remain equal.

D. Taxonomy Analysis (Entity Embedding and Euclidean Distance)

Taxonomic analysis relies on the findings outlined in reference [3], which indicate that certain virus families exhibit unique and consistent association patterns with specific insect families. For instance, literature from the past indicates that viruses belonging to the Geminiviridae family consistently have an insect vector from the Aleyrodidae family. Given that the Pepper Yellow Leaf Curl Virus also belongs to the Geminiviridae family, its potential insect vector could likewise be from the Aleyrodidae family.

The idea of taxonomic analysis is to take a list of candidate insects and analyze it by comparing their taxonomic distance and how similar they are to the relevant taxonomy. Relevant taxonomy refers to all taxonomic information (order, family, or genus) about insects that are generally known to be vectors of the virus (order, family, or genus) in some literature in the past. At this time, the user can input the relevant taxonomy information, or it can be prefilled by the system based on our association database. Fig. 9 shows steps in taxonomic analysis.

Firstly, the taxonomic data for each insect and relevant taxonomy needs to be converted into RDF Graph. After that the RDF graph is converted into numeric form using an entity embedding technique called RDF2Vec [37, 38]. Then, Euclidean distance (ED) [26] is used to measure the taxonomic similarity of each insect in that (numeric) vector space. The ED formula can be seen in Equation 2. The values of p and q are two points in the vector space. The qi and pi are vector values. n is the number of dimensions.

$$ED(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (2)$$

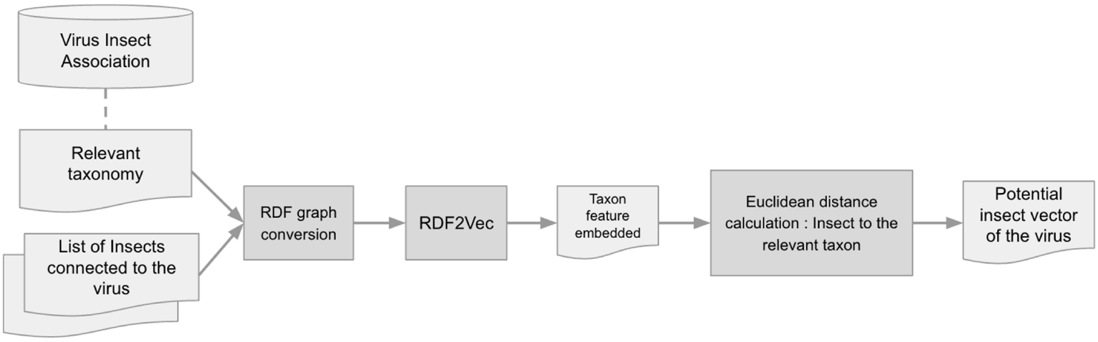

Virus Insect Association
Relevant taxonomy
List of Insects connected to the virus
RDF graph conversion
RDF2Vec
Taxon feature embedded
Euclidean distance calculation : Insect to the relevant taxon
Potential insect vector of the virus

Fig. 9 The taxonomy analysis steps

E. Final Score

The final score combines CM and ED using Equation 3. CM shows how important nodes are compared to other nodes. It aims to recognize nodes that are important or central among the whole list of other nodes in a graph. In contrast, ED shows how far the candidate insect vector point is from the relevant taxonomy in the insect taxonomic feature space. The greater the CM, the greater the final value (directly proportional), and the smaller the ED distance, the greater the final value (inversely proportional), resulting in Equation 3. A min-max scaler is used to normalize CM and ED, ensuring both have the same range of values between 0 and 1. The number 1 is also added to preserve the values if some value is zero.

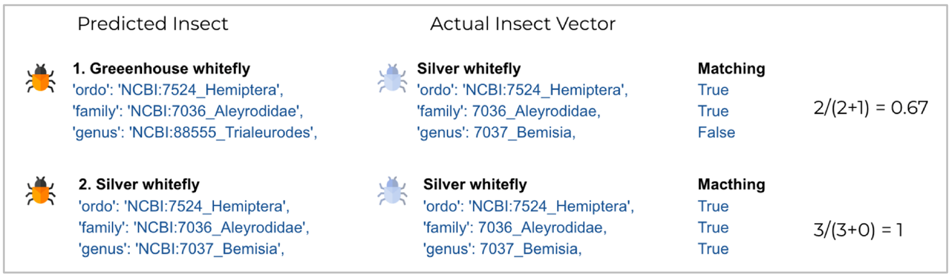
$$Final\ Score = \frac{1 + minMaxScale(DC)}{1 + minMaxScale(ED)} \qquad (3)$$

F. Testing

The testing stage aims to validate network analysis methods. Test data, consisting of 21 pairs of viruses and their confirmed insect vectors from [3, 39–41], can be accessed at https://linktr.ee/vektorpedia. The precision metric [42] is used to assess the accuracy of predicting the taxon paths of the actual insect vectors. The precision formula can be seen in Equation 4.

$$Precision = \frac{True\ Positive\ Taxon}{True\ Positive\ Taxon + False\ Positive\ Taxon} \qquad (4)$$

Since the network analysis returns the predicted insects in the form of a list, various testing options are available. The approach to testing can involve selecting a single insect with the highest final score, the top three, the top five, or even more, depending on specific requirements and the desired number of recommended insect vectors. This range of insect numbers is referred to as a prediction range. The testing concept is illustrated in Fig. 10.



Predicted Insect	Actual Insect Vector		
1. Greeenhouse whitefly 'ordo': 'NCBI:7524_Hemiptera', 'family': 'NCBI:7036_Aleyrodidae', 'genus': 'NCBI:88555_Trialeurodes',	**Silver whitefly** 'ordo': 'NCBI:7524_Hemiptera', 'family': 7036_Aleyrodidae, 'genus': 7037_Bemisia,	**Matching** True True False	2/(2+1) = 0.67
2. Silver whitefly 'ordo': 'NCBI:7524_Hemiptera', 'family': 'NCBI:7036_Aleyrodidae', 'genus': 'NCBI:7037_Bemisia',	**Silver whitefly** 'ordo': 'NCBI:7524_Hemiptera', 'family': 7036_Aleyrodidae, 'genus': 7037_Bemisia,	**Macthing** True True True	3/(3+0) = 1

Fig. 10 Proportion of (a) nodes based on their entity type; (b) edges based on interaction type

Fig. 10 depicts the predicted insect within the 2-prediction range. Assuming the actual insect vector is the Silver Whitefly, the precision is 0.67 for the first insect and 1.0 for the second insect. To represent this "more than one prediction range" test case, the prediction with the highest precision is selected, which, in this case, is the second insect. This indicates that the network analysis with a 2-prediction range has a precision of 1.0.

The decision not to use the average is justified by the fact that the average does not effectively represent the prediction results in this context. Another insect that does not fulfill the same taxon as the actual insect is not necessarily the wrong prediction and could also be a vector as long as they are in the same family. To accommodate this condition, the test focuses solely on determining the presence of actual insects in the range of predicted insects. As long as actual insects are identified among the predicted list, it is assumed that the predictions are accurate. However, to showcase the differences in results across various prediction ranges, the test cases are implemented for tolerance ranges of 1, 3, 5, and 10 prediction ranges.

G. Knowledge Extension

This stage aims to provide detailed information about the predicted insects. This stage is a complementary process where predicted insects are enriched with detailed information from knowledge bases such as Wikidata, DBPdia, and NCBI Taxon Ontology. The information provided from the data sources can be seen in Table 3.

TABLE 3
KNOWLEDGE EXTENSION DATA SOURCES

No	Data source	Information
1	Wikidata	Photos, scientific names, related literature and articles
2	DBPedia	Short description
3	GloBI	Interactions of insects with plants, viruses and biocontrol agents
4	NCBI Taxon Ontology	Taxonomic data of whole organisms such as superkingdom-species, and insect vector relatives

H. Application Development

This stage encompasses the development of a vector search engine application using a web platform. The application incorporates the analysis techniques previously described, and the resulting information is presented on a dashboard. There are three services to be built, namely: (1) a web client for the interface; (2) a computing server for the analysis and computation services; and (3) an NCBI Taxon Ontology Server to serve the data through SPARQL endpoint.

IV. RESULTS

To demonstrate how a research step works, the *Pepper Yellow Leaf Curl Virus* is selected as an example. The other viruses are examined in the testing subsection. The information from reference [3] is used as a benchmark. It says that the *Pepper Yellow Leaf Curl Virus* has an insect vector from the Aleyrodidae insect family, and the actual insect vector is *Bemisia Tabaci*.

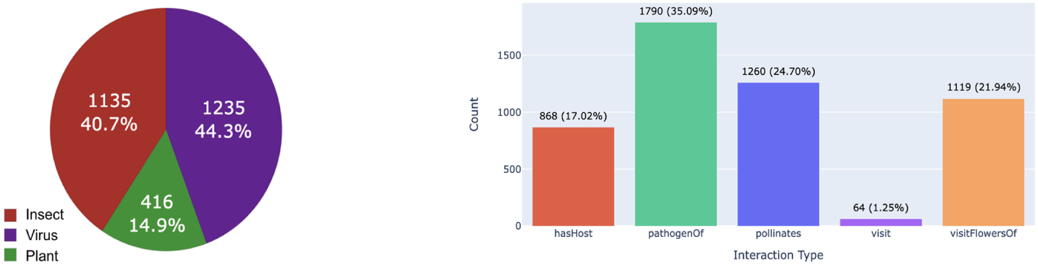
A. Data Acquisition

Table 4 shows changes in the amount of data obtained at each data acquisition sub-stage. Even though there were at most 48,624 rows of interaction data at one of the sub-stages, in the end, only 5,998 interactions (edges) were used. This is because some duplicate data is removed in the next sub-stage. In this way, even though the data is reduced, it is hoped that the interaction data obtained is truly of high quality.

Fig. 11 (a) shows the proportion of entities (nodes) in the data. The graph's most prevalent entities are viruses (44.3%), then insects (40.7%), and finally plants (14.9%). Fig. 11 (b) shows the proportion of interaction types (edges) in the data. In terms of the total number of interactions, *pathogenOf* (35.09%) was the most prevalent type of interaction, followed by *pollinates* (24.70%), *visitFlowerOf* (21.94%), *hasHost* (17.02%), and *visit* (1.25%).

TABLE 4
CHANGES IN DATA ON DATA ACQUISITION

Sub-stage	Amount
1. Virus Interaction (Depth I):	
1.1. Data retrieval	1028 interaction
1.2. Table Splitting	516 node, 516 edge
1.3. Disambiguation	514 node, 511 edge
2. Plant and Insect Interaction (Depth II):	
2.1. Data retrieval	48624 Interaction
2.2. Table splitting	7369 node, 12409 edge
2.3. Disambiguation	6874 node, 11231 edge
2.4. Table concatenation	7388 node, 11742 edge
2.5. Final result (after filter)	2786 node, 5998 edge



(a) (b)

Fig. 11 Proportion of (a) nodes based on their entity type; (b) edges based on interaction type

B. Graph Conversion

Fig. 12 shows the visualization of the generated graph from the node and edge tables. Several red nodes (insects) occur in the center of the graph. The position in the center indicates that the node has many edges (interactions). This also indicates that those nodes have the potential to become insect vectors. After manual inspection, it was discovered that Bemisia Tabaci (the vector of the Pepper Yellow Leaf Curl Virus) was present in the middle of the graph. Interaction analysis can be used to obtain these centered insect nodes automatically.

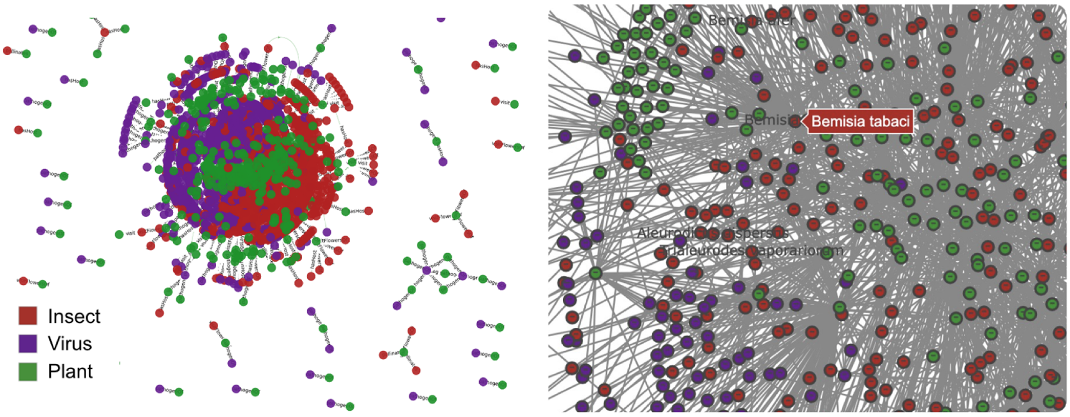


Fig. 12 Data conversion from tabular to Graph

C. Interaction Analysis

In this section, the results of interaction analysis when using original DC and TNC-DC are compared. Fig. 13 shows the analysis results when using the original DC. Fig. 15 shows the analysis results when using TNC-DC. The x-axis is the degree value, and the y-axis is the name of the insect.

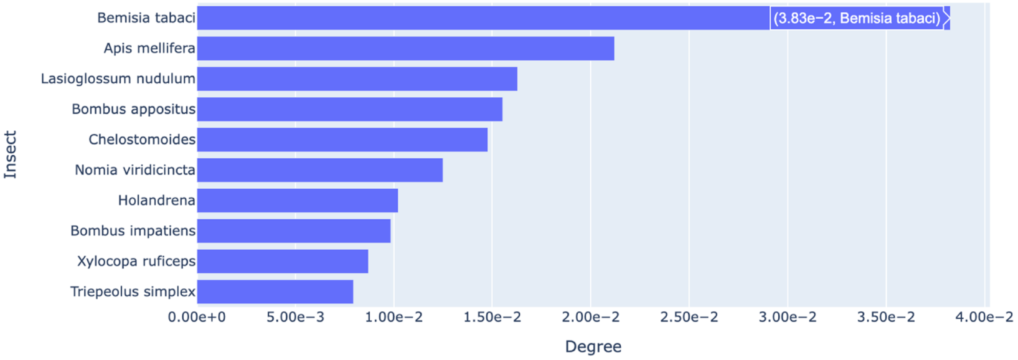
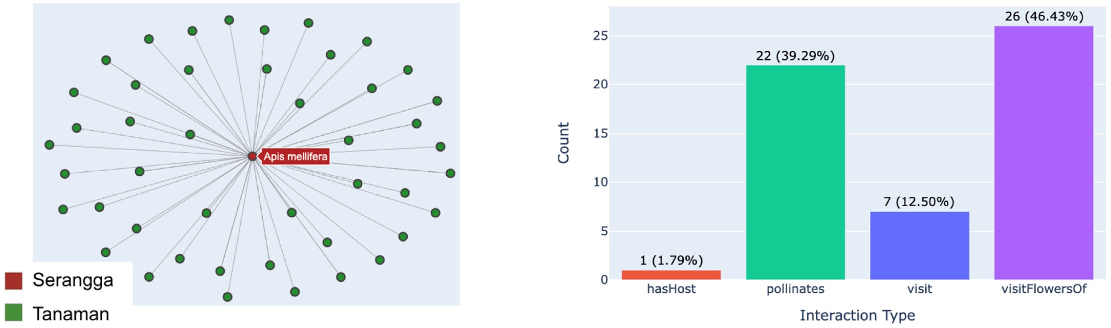


Fig. 13 Some results from original DC

Fig. 13 shows that Bemisia Tabaci was placed first in the original DC results. However, this was followed by various species of bees, namely Apis Mellifera, Lasioglossum Nudulum, Bombus Appositus, Chelostomoides, Nomia Viridicincta, Holandrena, Bombus Impatiens, Xylocopa Ruficeps, Triepeolus Simplex. In fact, each bee has no interaction at all with the virus, and it is not a vector for the Begomovirus virus. Fig. 14 shows the neighbor nodes that Apis Mellifera bees have and their interaction types.



(a) (b)

Fig. 14 Apis Mellifera's (a bee) neighbor nodes (a) and its interaction types (b)

In Fig. 14 (a), it can be seen that although Apis Mellifera has a high degree of interaction, its interactions are only with plants. Apart from that, in Fig. 14 (b), it can be seen that the Apis Mellifera interaction type is only between hasHost, pollinates, visit and visitFlowersOf. This indicates that the reason why bees have a high degree is because they interact with many plants. However, their interaction is only about pollination events and not virus transmission events. There are concerns that bees at the highest degree hide other insects that actually interact with the virus. Therefore, we modified the original DC to become TNC-DC.

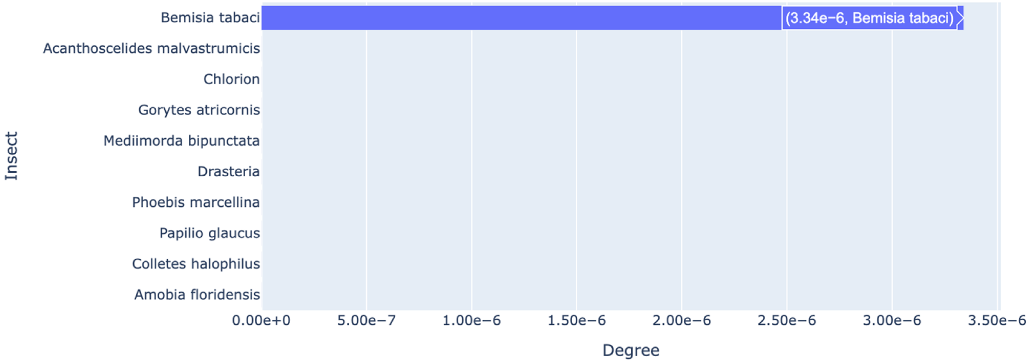
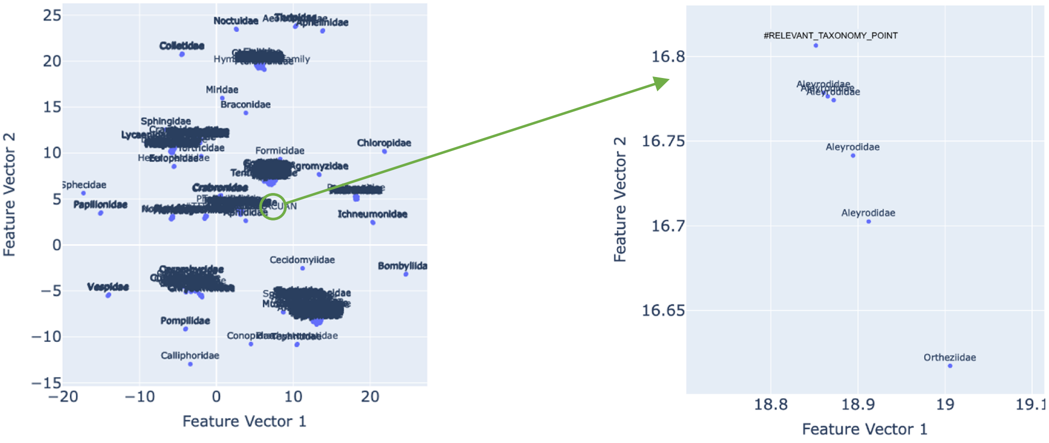


Fig. 15 Some results from TNC-DC

In Fig. 15, it can be seen that TNC-DC only produces Bemisia Tabaci in the first place, while the other insects have a degree value of 0. This low degree value is caused by the insect not having any interaction with the virus. Some bees that were previously at the highest places in the original DC now have equivalent value to other insects that also have no interaction with the virus. Even though most of the TNC-DC results are 0, all those insects can still be analyzed for their potential based on taxonomy at a later stage.

D. Taxonomy Analysis (Entity Embedding and Euclidean Distance)

Fig. 16 illustrates the results of embedding insect data that has been reduced to two dimensions. The label at each dot indicates the insect family. Fig. 16 (a) shows that each insect has been gathered based on taxonomy. In Fig. 16 (b), since the *Pepper Yellow Leaf Curl Virus* insect vector comes from the Aleyrodidae family, a closer examination is performed within the Aleyrodidae group. It can be seen that five insects belong to the Aleyrodidae taxonomy. Then, Euclidean distance is employed to find these five insects. Fig. 17 shows the results of the Euclidean distance. The shorter the distance, the more potential insects there are as vectors. The top five insects are Trialeurodes Vaporariorum (0.271), *Bemisia Tabaci* (0.298), Aleurodicus Dispersus (0.317), Bemisia Emiliae (0.320), and Bemisia Afer (0.457).



(a)

(b)

Fig. 16 (a) Embedded data; (b) Zoom in to Aleyrodidae family

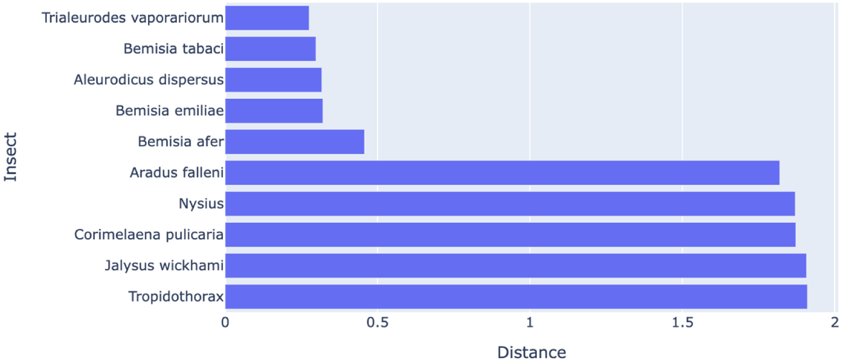


Fig. 17 Some results of Euclidean distance calculation

E. Final Score

Fig. 18 shows the results of calculating the final score. The higher the final score, the greater the insect's potential as a vector. Since the final score is a combination of network analysis and taxonomy analysis, the insects are now listed in a different order: Bemisia Tabaci (1.0), Trialeurodes Vaporariorum (0.5), Aleurodicus Dispersus (0.49), Bemisia Emiliae (0.49), and Bemisia Afer (0.48). Based on these results, *Bemisia Tabaci* is a virus vector of the *Pepper Yellow Leaf Curl Virus*. This result has been validated by our benchmark data [3]. At a later stage, information about *Bemisia Tabaci* is enriched.

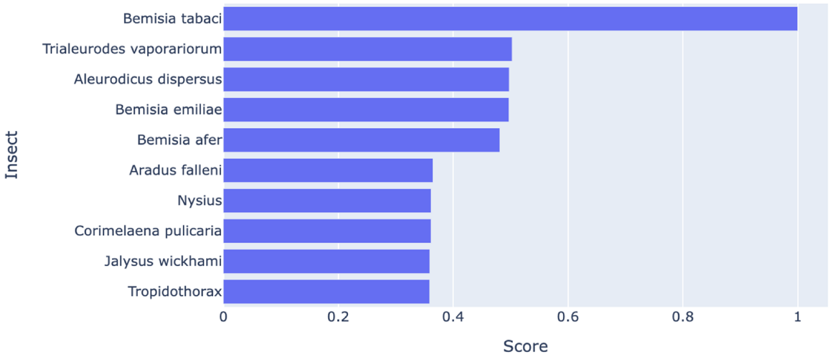


Fig. 18 Some results of final score calculation

F. Testing

The test results are available at https://linktr.ee/vektorpedia. Fig. 19 shows a comparison of precision values for each prediction range. The y axis represents average precision, and the x axis represents prediction range.

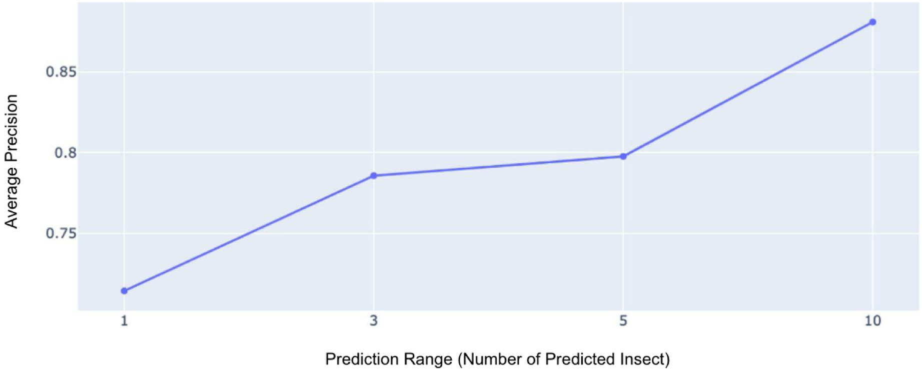


Fig. 19 Precision values in every prediction range

From Fig. 19, it can be seen that there is an increase in average precision when the prediction range is expanded. A precision of 0.71 was obtained within the 1-prediction range, 0.78 within the 3-prediction range, 0.8 within the 5-prediction range, and 0.88 within the 10-prediction range. The precision score indicates that the analysis method can effectively predict the actual insect vector, particularly starting from the 5-prediction range. It implies that the actual insect vector may be ranked within the first to fifth positions in the list resulting from network analysis. Considering the recommendation of the smallest number of insects with the highest precision, the top five insects can be recommended to the user.

G. Knowledge Extension

The enhancement stage generates detailed information regarding the vector insect resulting from the analysis, *Bemisia Tabaci*. Information about the pathogens (natural enemies) of *Bemisia Tabaci* was discovered when *Bemisia Tabaci* was integrated into GloBI. Fig. 20 shows the natural enemies of the *Bemisia Tabaci* and their proportions. There are 63 insects, 18 viruses, and 41 other organisms (arthropods, fungi, etc.) that are natural enemies or parasites of *Bemisia Tabaci*. The complete list of *Bemisia Tabaci's* natural enemies can be accessed at https://linktr.ee/vektorpedia. There are Eretmocerus, Encarsia, and Aphelinus, which are natural enemies and can be used as biocontrol for *Bemisia Tabaci*.

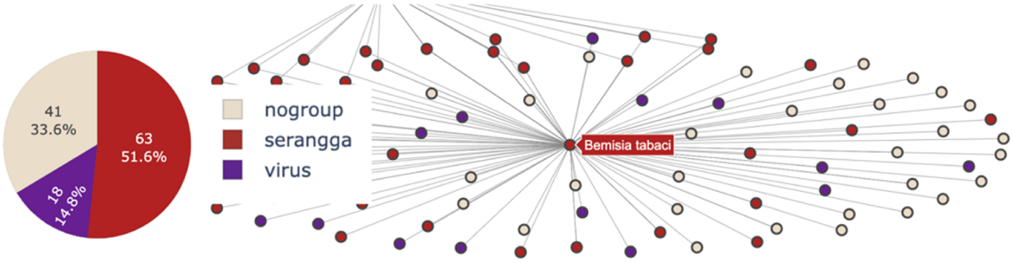


Fig. 20 Bemisia Tabaci's natural enemy graph

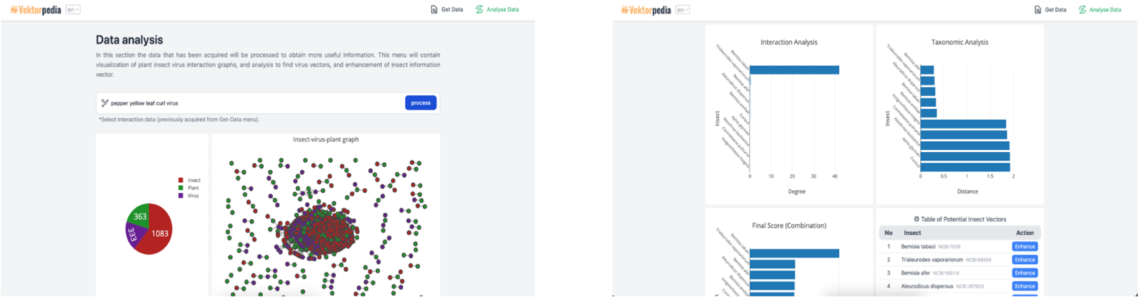
Relatives of the *Bemisia Tabaci* insect from another species were identified in the NCBI taxon ontology. Insect relatives of *Bemisia Tabaci* can be accessed at https://linktr.ee/vektorpedia. There are 37 different insect relatives of *Bemisia Tabaci*. Detailed information about the *Bemisia Tabaci*, including common names, photographs, and descriptions, was found in Wikidata and DBPedia. Table 5 shows detailed information on *Bemisia Tabaci*.

TABLE 5
DETAILED INFORMATION FROM WIKIDATA AND DBPEDIA

No	Proces	Information
1	Scientific Name	Bemisia Tabaci
2	Common Names	Sweet potato whitefly, etelänjauhiainen, bomullsmellus, 煙草粉蝨種群,煙草粉蝨,銀葉粉蝨, عثة التبغ, Tabakmottenschildlaus, Aleurode du tabac, シルバーリーフコナジラミ, 담배가루이, Tabakswittevlieg, Mosca-branca (Bemisia tabaci), dan Табачная белокрылка.
3	Photo	https://commons.wikimedia.org/wiki/File:Silverleaf_whitefly.jpg
4	Description	The silverleaf whitefly (Bemisia Tabaci, also informally referred to as the sweet potato whitefly) is one of several species of whitefly that are currently important agricultural pests...

H. Application Development

The application development phase has been successfully carried out and resulted in a web-based vector search engine application. Fig. 21 (a) is a screenshot of the graph visualization stage. Fig. 21 (b) is a screenshot of the network analysis stage. The application can be accessed via https://ipb.link/vektorpedia.



Data analysis

In this section the data that has been acquired will be processed to obtain more useful information. This menu will contain visualization of plant insect virus interaction graphs, and analysis to find virus vectors, and enhancement of insect information vector.

pepper yellow leaf curl virus

*Select interaction data (previously acquired from Get-Data menu).

Table of Potential Insect Vectors

No	Insect	Action
1	Bemisia tabaci NCBI:7038	Enhance
2	Trialeurodes vaporariorum NCBI:88556	Enhance
3	Bemisia afer NCBI:166114	Enhance
4	Aleurodicus dispersus NCBI:267823	Enhance

(a)

(b)

Fig. 21 (a) Graph visualization stage in application; (b) Network analysis stage in application

V. DISCUSSION

Identifying insect vectors is crucial for breaking transmission and mitigating the negative impact on agricultural production. However direct insect vector identification in the field poses several challenges. Our study provides some alternative approaches to identify insect vectors indirectly by analysing a big biodiversity interaction data provided by GloBI [16]. This idea led to the development of an insect vector search engine system named "Vektorpedia". By implementing network analysis and taxonomy analysis, Vektorpedia effectively aids in finding insect vectors of some viruses and provides the useful information such as the natural enemies of the insect vector, the other varian and species of the insect vector, etc. This is an important step towards increasing agricultural productivity and sustainability. This stage of research drew inspiration from a previous study, as mentioned in reference [43]. However, a notable distinction lies in the objective that [43] primarily centered around COVID-19 information, whereas our investigation was specifically oriented towards insect vectors.

For comparison, there is another study [44] that also uses analytical data for studies on insect-plant-virus interactions, but this research uses proteomic techniques. However, research on proteomic techniques has several gaps that can be covered by our research. Proteomic techniques used in this study may not capture the full complexity of the interactions, as they primarily focus on protein identification and quantification. This may not immediately provide detailed information about the natural enemies of insect vectors. Since our study uses a biotic interaction approach, obtaining information about natural enemies becomes more straightforward. In addition, the simple interaction analysis offers increased accessibility and cost-effectiveness, eliminating the need for specialized equipment or expertise in data collection and management. Furthermore, simple interaction analysis saves more time when compared to traditional or direct insect vector identification [8, 9]. Which requires preparation before entering the field and also requires special tools and skills.

However, our study also has several limitations. It is limited by the availability of data on GloBI. Some virus interaction data may not yet be available. In addition, an understanding of the common virus names is required to access and understand the results. Apart from that, some actual insect vectors are not successfully sorted in the first place. Some actual vector insects are actually in the first to fifth place in order by final score. This is because the taxonomy analysis does not differ significantly in distance for insects in the same family. Due to the dynamic arrangement of the insect vectors, multiple experiments in the range of predicted insects were conducted. Instead of using one insect as predicted insect, we immediately obtained good precision values in the third, fifth, and tenth prediction ranges of predicted insects. It was discovered that raising the prediction range would also raise the precision. It means that the actual vector is actually obtained, but it's just a matter of ordering. The testing results show that our analysis approach has a precision value of 0.78 in third prediction range, 0.80 in fifth prediction range, and 0.88 in tenth prediction range. Additionally, even though they are not actual insects, the insects that frequently change their arrangement still belong to the same family, so they still have the potential to be vectors; it's just a matter of the test assessment, which needs to be compared to actual insects.

For future work, we recommend trying to use RDF data and perform analysis directly via SPARQL queries rather than our data acquisition step, which requires data conversion and analysis, which may require more execution time. This is also related to the [44] study, which states that querying graph data is faster than tabular relational data. Apart from that, we also recommend conducting research to uncover more information on taxonomic association patterns of insects and viruses and documenting them into knowledge graphs or ontologies.

VI. CONCLUSIONS

In conclusion, our study introduces the "Vektorpedia" insect vector search engine, utilizing biodiversity interaction data from GloBI. Through network and taxonomy analyses, Vektorpedia offers an indirect yet effective approach to identify insect vectors, providing valuable insights into natural enemies, insect variations, and species. This innovative method proves to be accessible, cost-effective, and time-efficient compared to traditional techniques. Despite limitations in data availability and taxonomy analysis, our precision values demonstrate the reliability of our approach. For future research, we propose exploring RDF data and SPARQL queries for direct analysis, aiming to enhance efficiency. Additionally, further investigation into taxonomic association patterns between insects and viruses can contribute to knowledge graphs or ontologies, advancing our understanding of these crucial interactions.

Author Contributions: *Moh. Zulkifli Katili*: Conceptualization, Methodology, Writing - Original Draft, Editing, Data Collection, Software. *Yeni Herdiyeni*: Conceptualization, Methodology, Supervision, Review. *Medria Kusuma Dewi Hardhienata*: Conceptualization, Methodology, Supervision, Review.

All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive a specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability: The data and source code used in this paper are publicly available to promote the transparency and reproducibility of this paper's findings. The insect vector search engine web can be accessed here: https://ipb.link/vektorpedia. Access the experiment and data analysis here: https://github.com/kikiraihan/plant-virus-vector. This link provides a Python notebook and the backend code of the application. The test data can be accessed via the following link: https://ipb.link/vektorpedia-example-data. Additionally, we also provide access to the front-end source code of the vector insect search engine application. It can be accessed via the following link:

https://github.com/kikiraihan/vue-thesis. Several table in result section can be accessed at https://linktr.ee/vektorpedia.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Katili: https://orcid.org/0000-0001-7067-1705
Herdiyeni: https://orcid.org/0000-0002-3389-1730
Hardhienata: https://orcid.org/0000-0002-7646-1896

REFERENCES

[1] A. E. Whitfield, B. W. Falk, and D. Rotenberg, "Insect vector-mediated transmission of plant viruses," in *Virology*, vol. 479–480, pp. 278–289, 2015, doi: https://doi.org/10.1016/j.virol.2015.03.026.

[2] M. Deshoux, B. Monsion, and M. Uzest, "Insect cuticular proteins and their role in transmission of phytoviruses," in *Curr. Opin. Virol.*, vol. 33, pp. 137–143, 2018, doi: https://doi.org/10.1016/j.coviro.2018.07.015.

[3] M. A. Catto, H. Mugerwa, B. K. Myers, S. Pandey, B. Dutta, and R. Srinivasan, "A review on transcriptional responses of interactions between insect vectors and plant viruses," in *Cells*, vol. 11, no. 4, p. 693, 2022.

[4] Y. S. Pauzi, "Deteksi Garlic common latent virus dan Shallot latent virus pada Beberapa Fase Pertumbuhan Tanaman Bawang Merah dan Bawang Putih [skripsi]," Bogor: Institut Pertanian Bogor, 2017.

[5] Meliyana, "Virus Infection and Growth Performance of Several Local Garlic Cultivars [skripsi]," Bogor: Institut Pertanian Bogor, 2020.

[6] Y. M. A. Sandra, "Penapisan Dan Identifikasi Karakter Ketahanan Terhadap Virus Gemini Dan Kutu Kebul Pada Cabai [tesis]," Bogor: Institut Pertanian Bogor, 2022.

[7] D. Wahyudin, "Penapisan Ketahanan Galur dan Taksasi Kehilangan Hasil Tomat (Lycopersicum esculentum L.) Terhadap infeksi Tomato chlorosis crinivirus [tesis]," Bogor: Institut Pertanian Bogor, 2022.

[8] C. Lilies, "Kunci determinasi serangga," in *Kanisius. Yogyakarta*, vol. 223, 1991.

[9] M. Heviyanti and M. Syahril, "Identifikasi serangga hama tanaman padi sawah (Oryza sativa L.) di Desa Paya Rahat, Kecamatan Banda Mulia, Kabupaten Aceh Tamiang," in *Prosiding Seminar Nasional Pertanian*, 2018.

[10] H. Lee *et al.*, "Insect vector manipulation by a plant virus and simulation modeling of its potential impact on crop infection," in *Sci. Rep.*, vol. 12, no. 1, p. 8429, 2022.

[11] L. Rubio, L. Galipienso, and I. Ferriol, "Detection of plant viruses and disease management: Relevance of genetic diversity and evolution," in *Front. Plant Sci.*, vol. 11, p. 1092, 2020.

[12] S. dan N. Yasin, "Karakterisasi Kutu Kebul (Bemisia Tabaci) Sebagai Vektor Virus Gemini Dengan Teknik Pcr-rapd," in *J. Hama dan Penyakit Tumbuh. Trop.*, vol. 6, no. 2, pp. 113–119, 2006, doi: 10.23960/j.hptt.26113-119.

[13] L. Xiaoxue, B. Xuesong, W. Longhe, R. Bingyuan, L. Shuhan, and L. Lin, "Review and trend analysis of knowledge graphs for crop pest and diseases," in *IEEE Access*, vol. 7, pp. 62251–62264, 2019.

[14] D. Wood, M. Zaidman, L. Ruth, and M. Hausenblas, *Linked Data : Structured data on the web*. New York, NY, USA: Manning Publications Co., 2014.

[15] A. L. Barabasi, M. Tillich, K. Albrecht, M. Martino, M. Pósfa, and G. Musella, *Network Science Book*. Cambridge: Cambridge University Press, 2015. [Online]. Available: http://barabasi.com/networksciencebook/

[16] J. H. Poelen, J. D. Simons, and C. J. Mungall, "Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets," in *Ecol. Inform.*, vol. 24, pp. 148–159, 2014.

[17] M. M. Yusof, N. F. Rosli, M. Othman, R. Mohamed, and M. H. A. Abdullah, "M-DCocoa: M-agriculture expert system for diagnosing cocoa plant diseases," in *Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, February 06-07, 2018*, Springer, 2018, pp. 363–371.

[18] W. Jearanaiwongkul, C. Anutariya, and F. Andres, "An ontology-based approach to plant disease identification system," in *Proceedings of the 10th International Conference on Advances in Information Technology*, 2018, pp. 1–8.

[19] K. Lagos-Ortiz, M. D. Salas-Zárate, M. A. Paredes-Valverde, J. A. García-Díaz, and R. Valencia-García, "AgriEnt: A Knowledge-Based Web Platform for Managing Insect Pests of Field Crops," *Applied Sciences*, vol. 10, no. 3. 2020. doi: 10.3390/app10031040.

[20] W. Jearanaiwongkul, C. Anutariya, T. Racharak, and F. Andres, "An Ontology-Based Expert System for Rice Disease Identification and Control Recommendation," in *Appl. Sci.*, vol. 11, no. 21, p. 10450, 2021.

[21] M. Á. Rodriguez-Garcia, F. García-Sánchez, and R. Valencia-García, "Knowledge-Based System for Crop Pests and Diseases Recognition," *Electronics*, vol. 10, no. 8. 2021. doi: 10.3390/electronics10080905.

[22] K. A. Garrett *et al.*, "Network analysis: a systems framework to address grand challenges in plant pathology," in *Annu. Rev. Phytopathol.*, vol. 56, pp. 559–580, 2018.

[23] A. Bundy and L. Wallen, "Breadth-first search," in *Cat. Artif. Intell. tools*, p. 13, 1984.

[24] L. C. Freeman, "Centrality in social networks conceptual clarification," in *Soc. Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[25] B. Steenwinckel, G. Vandewiele, T. Agozzino, and F. Ongenae, "pyRDF2Vec: A Python Implementation and Extension of RDF2Vec," in *European Semantic Web Conference*, Springer, 2023, pp. 471–483.

[26] P.-E. Danielsson, "Euclidean distance mapping," in *Comput. Graph. image Process.*, vol. 14, no. 3, pp. 227–248, 1980.

[27] J. Lehmann *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," in *Semant. Web*, vol. 6, no. 2, pp. 167–195, 2015.

[28] M. Farda-Sarbas and C. Mueller-Birn, "Wikidata from a Research Perspective--A Systematic Mapping Study of Wikidata," in *arXiv Prepr. arXiv1908.11153*, 2019.

[29] S. Schulz, H. Stenzhorn, and M. Boeker, "The ontology of biological taxa," in *Bioinformatics*, vol. 24, no. 13, pp. i313–i321, 2008.

[30] E. Arnaud *et al.*, "The ontologies community of practice: A CGIAR initiative for big data in agrifood systems," in *Patterns*, vol. 1, no. 7, p. 100105, 2020.

[31] A. Hogan *et al.*, "Knowledge graphs," in *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, 2021.

[32] K. Seltmann, J. Poelen, K. Sullivan, and J. Zaspel, "Making Parasite-Host Associations Visible using Global Biotic Interactions," in *Biodivers. Inf. Sci. Stand.*, vol. 4, 2020.

[33] Y. Do and M. B. Choi, "Network analysis for co-occurrence of pest insects on host crops," in *Entomol. Res.*, vol. 49, no. 1, pp. 35–45, 2019.

[34] E. Delmas *et al.*, "Analysing ecological networks of species interactions," in *Biol. Rev.*, vol. 94, no. 1, pp. 16–36, 2019.

[35] T. Strydom *et al.*, "A roadmap towards predicting species interaction networks (across space and time)," in *Philos. Trans. R. Soc. B*, vol. 376, no. 1837, p. 20210063, 2021.

[36] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using NetworkX," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[37] P. Ristoski and H. Paulheim, "Rdf2vec: Rdf graph embeddings for data mining," in *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15*, Springer, 2016, pp. 498–514.

[38] G. Vandewiele, B. Steenwinckel, T. Agozzino, and F. Ongenae, "pyRDF2Vec: A Python Implementation and Extension of RDF2Vec," in *arXiv Prepr. arXiv2205.02283*, 2022.

[39] G. M. Angelella and T. D. Waters, "Afidopyropen as a potential tool for Potato leafroll virus management in post-neonicotinoid potato production," in *J. Econ. Entomol.*, vol. 116, no. 3, pp. 713–718, 2023, doi: https://doi.org/10.1093/jee/toad042.

[40] R. Thangjam, V. Kadam, P. D. Nath, and R. K. Borah, "Insect Vectors Associated with Viral Diseases of King Chilli (Capsicum Chinense Jacq.) in North East India," in *Indian J. Entomol.*, pp. 1–4, 2022.

[41] M. F. M. Santiago, K. C. King, and G. C. Drew, "Interactions between insect vectors and plant pathogens span the parasitism–mutualism continuum," in *Biol. Lett.*, vol. 19, no. 3, p. 20220453, 2023.

[42] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*, Springer, 2005, pp. 345–359.

[43] F. Darari, "COVIWD: COVID-19 Wikidata Dashboard," in *J. Ilmu Komput. dan Inf.*, vol. 14, no. 1, pp. 39–47, 2021.

[44] P. Mittapelly and S. P. Rajarapu, "Applications of Proteomic Tools to Study Insect Vector–Plant Virus Interactions," *Life*, vol. 10, no. 8. 2020. doi: 10.3390/life10080143.

Publisher's Note: Publisher stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.