PythoPharmVis: A Network Analysis tool to Identify Key Entities and Communities for Phytomedicine and Pharmacogenomics

Emmanuel Ednalan and Geoffrey Solano

Department of Physical Sciences and Mathematics,
College of Arts and Sciences,
University of the Philippines Manila, Ermita, Manila, PH
edednalan@up.edu.ph, gasolano@up.edu.ph

Abstract. Network science offers a powerful way of exploring and discovering on knowledge and insights. Making use of graphs as mathematical structures, it is able to model pair-wise relations between objects, thus, providing in-depth understanding of relationships of entities within and among networks and structures across various disciplines. The motivation of this study is to explore how it can be useful in the study of the relationships between plants, small-molecule drugs, genes and diseases, particularly in the identification of key entities and detection of communities. Data from open-source databases were used to derive the association networks, particularly, Dr. Duke's Phytochemical and Ethnobotanical Databases for the plants and drugs datasets; BioGrid for the drugs and genes datasets; and PharmKB for the drugs and genes, as well as genes and diseases datasets. The networks were analyzes using the weighted degree centrality and the Leiden algorithm for community detection. A web application was also developed, which allows the users to view and analyze subnetworks based on certain plants, smds, genes or diseases of interest. It is hoped that the results of this study as well as the tool developed will help in phytomedical and pharmacogenomic research.

Keywords: plants, phytochemicals, small-molecule drugs, genes, diseases, centrality, community detection, graph database

1 INTRODUCTION

Plants have been used for centuries for their therapeutic and healing capabilities. They have a wide range of biological activities which are useful for the prevention and treatment of diseases [6]. Plants contain phytochemicals which can be small-molecule bioactive compounds which they use for protecting themselves from natural risks. These phytochemicals can be extracted from plants as they may exhibit pharmacological effects in the human body such as antimicrobial, antiallergic, and more. Thus, plants and their phytochemicals continue to be considered as alternatives to synthetic drugs, and phytomedicine as a promising research area.

J. Caro et al. (eds.), Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2024) , Atlantis Highlights in Computer Sciences 23,
https://doi.org/10.2991/978-94-6463-684-0_15

Network science offers a powerful way of exploring, discovering, and researching on biomedical knowledge [3]. It can be used to help in having in-depth understanding of relationships of entities within and among networks, structures, importance, and more. Making use of graphs as mathematical structures, it is able to model pair-wise relations between objects. Moreover, graphs consist of nodes and edges, forming structures such as a graph database for designing a database using graph structures, and graph based representation, which gives different points of view and make problems easier to solve [2]. Networks have a variety of applications and can be implemented in almost every field of study.

One important task in network analysis is the determination of the importance or influence of a node within a network, which is known as centrality. Another important task is community detection or the determination of sets of nodes which are relatively similar or highly associated with each other. Existing methods make use of clustering as input that will be refined to form well-connected networks, composed of intra-community edges that connect nodes within a community and inter-community edges that connect two or more communities [1].

These capabilities of network analysis have allowed it to be useful in various domains such as sociology, epidemiology, business, even in healthcare and medicine. The motivation of this study is to explore how it can be useful in the study of the relationships between plants, small-molecule drugs(smds), genes and diseases, which shall henceforth be referred to in this study as biomedical entities. Health and phytomedicine researchers are constantly trying to discover and learn from different biomedical entities and their relationships as well as drug alternatives that are crucial to health, especially in the Philippines, with its rich biodiversity of flora, making it a ripe ground for research in phytomedicine and pharmacogenomics.

Furthermore, there is a need for a tool that would enable researchers in health and phytomedicine to study and and explore such relationships among the four biomedical entities that may help particularly in discovering and prescribing alternative treatments. There is an existing tool that provides these capabilities, however, it uses a different set of algorithms and does not use a graph database for faster graph retrieval. Additionally, it follows a different set of associations, as seen in Figure 1, where there are direct associations that can be adjusted and/or corrected. Research and inquiry with phytomedicine researchers have allowed the further refinement of modeling these relationships, in particular, that smds may target a protein that is produced by a certain set of genes and that some genes are associated with diseases. Figure 2 shows the modified entity-relationship diagram (ERD) among the four biomedical entities.

2 RELATED WORK

Recario et. al. [5] developed a database web application which allows users to contribute information, and is able to provide visualizations of the associations of the same four biomedical entities, namely: plants, smds, human conditions or

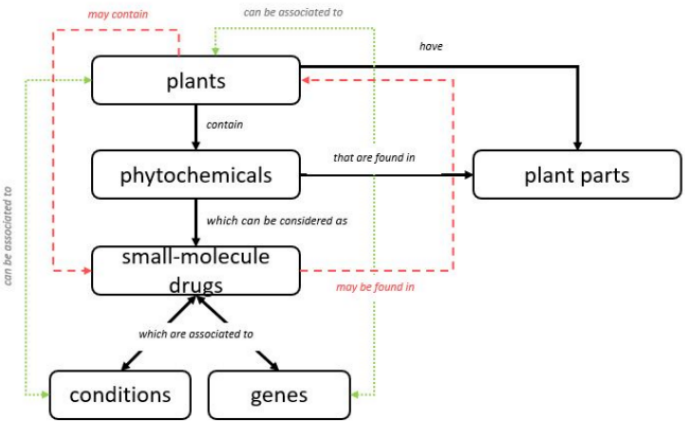

Fig. 1. Association of the four Biomedical entities. [5]

diseases, and human genes, in the form of graphs and tables. This study [5] did not use network-based clustering but instead made use of K-means clustering and Self-Organizing Maps for detection of communities.

The study of Wong et. al. [8] gives understanding and simplicity on the relationships between genes, drugs, and diseases, in an interdisciplinary manner, where pharmacogenomics meets the computing sciences. To identify these relationships, manual curation and automated data mining approaches were done on biomedical literature to makeup the datasets/databases to be used in the study. Consequently, GeneDive was developed as the product of this study, which is a web application for pharmacogenomics researchers and precision medicine practitioners that makes data, information, and visualization on the genes, diseases, drugs, and their interactions to be easily accessible, usable, and comprehensible through its search and visualization capabilities.

The work of Xiong et. al. [9] focuses on the interactions that are present between drugs. It is known that drug-drug interactions (DDI) may trigger adverse effects among patients and are a known threat to public health. Although there is an overabundance of biomedical information that involve DDI, only a few open-access knowledge systems or databases are available. The study [9] aims to implement an open-access online database for the interactions of drugs in different manners and help in facilitating clinical prescription and avoid dangerous drug combinations.

Groza et. al. [4] focused on repurposing of drugs with the assumption that different medicine have multiple functions, so alternatives are a possibility. It makes use of the ever-growing drug databases, which are accessible to researchers. They introduced a drug repurposing pipeline that is automated, which is based on interaction of drug-gene data, from DrugBank, and built a DG-interaction network. A bipartite graph composed of drug and gene interactions was constructed

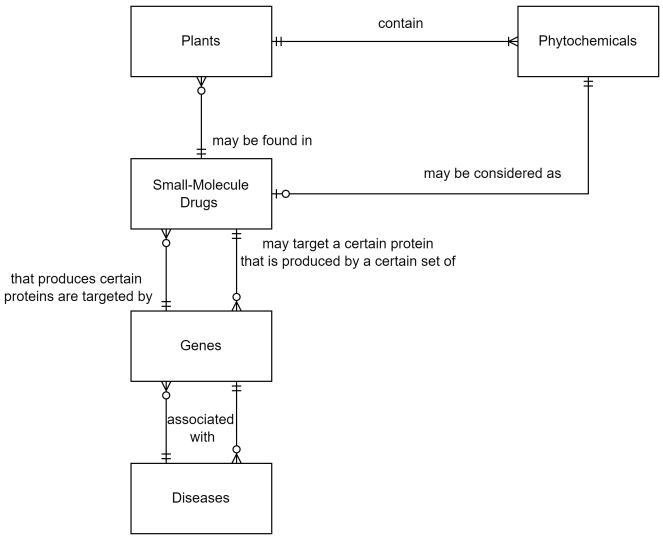

Fig. 2. Entity Relationship Diagram of the Biomedical Entities

and then used to create a weighted drug-drug similarity networks, where drugs with similar gene interactions form links between them. Drugs with high similarity can be further explored for repurposing through validation experiments.

3 METHODOLOGY

3.1 Dataset

PythoPharmVis uses databases that are not endemic in the Philippines. For the associations between plants and drugs, the datasets were downloaded from Dr. Duke's Phytochemical and Ethnobotanical Databases. For the associations between drugs and genes, as well as the associations between genes and diseases, the datasets were downloaded from BioGrid and PharmKB. All of these are in csv format, uploaded in an online database, and are used to create aggregated datasets for the algorithms to be used on.

3.2 Network Construction : Nodes, Edges and Weights

The set of networks that are available in the system have three types, namely the unweighted bipartite networks, weighted monopartite networks and the weighted bipartite networks. The unweighted bipartite networks can be easily derived using the direct links among the the entities in Figure 2 (phytochemicals not included) from the dataset, i.e. plants-smds, smds-genes and genes-diseases. Table 1 shows the number of the nodes and edges in the unweighted bipartite networks.

It should be noted that in the construction of the unweighted bipartite networks, say entity A to entity B, the focus is on entity A and the necessary connections with entity B are extracted. For instance, in the case of the plants to smds bipartite network, the focus is on plants and the necessary connections with smds from the different databases are extracted. It was observed that an smd connection found in one database may not be found in another, but if the connection is found in at least one database, it is included. It was observed that narrowing the condition such that a connection is included only if it is present in all databases greatly reduces the network, consequently reducing the information and insight that might be gained by researchers. This approach in network construction is also the reason why the number of nodes in the plants to smds bipartite network is not the same as the number of nodes in the smds to plants bipartite network.

Nodes and Edges (Unweighted Bipartite Networks)		
Data	Nodes	Edges
plant to smds	27147	85510
smd to genes	34951	14984
smd to plants	34057	85016
gene to smds	30881	14807
gene to diseases	31742	187745
disease to genes	5505	5485

Table 1. Unweighted Bipartite Networks Total Nodes and Edges

On the other hand, Table 2 is the overview of the nodes and edges of weighted monopartite networks derived from the bipartite graphs, wherein given a bipartite network with partitions A and B, a monopartite network A is formed from using the nodes in partition A, and the weight of two nodes in network A are the common neighbors they share in partition B. Similarly, a monopartite network B is formed from using the nodes in partition B, and the weight of two nodes in network A are the common neighbors they share in partition A.

Nodes and Edges (Weighted Monopartite Networks)		
Data	Nodes	Edges
plant to plant via smds	2376	740425
smd to smd via plants	31742	3983320
smd to smd via genes	31742	187745
gene to gene via smds	24239	81789
gene to gene via diseases	24239	66749
disease to disease via genes	3971	36897

Table 2. Weighted Monopartite Networks Total Nodes and Edges

The third type of network derived is shown in Table 3. The weighted bipartite networks are derived from two different unweighted bipartite networks, one with partition A and partition B, and the other with partition B and partition C. Clearly, both share a common partition, i.e. partition B. Thus, another bipartite network can be constructed with partition A and partition C, and the weights of the edges between a node in A and a node in C is the number of common neighbors they share in partition B. Hence we have the networks plant to gene via smds, and smds to diseases via genes.

Nodes and Edges (Weighted Bipartite Networks)		
Data	Nodes	Edges
plant to gene via smds	26615	48793
smds to diseases via genes	35713	109295

Table 3. Weighted Biartite Networks Total Nodes and Edges

3.3 Identification of Central Nodes and Communities

Weighted degree centrality measures and considers the degrees or number of edges a node has, and computes the weight per node in a community based on them. It is helpful in finding which are the most central entities, helping in formulating alternatives and in future researching in this field, since the interactions/relationships among the chosen four biomedical entities are crucial and important in today's life. It may also be used to determine which of these entities can be grouped together.

Clustering can be done via community-detection algorithm. One example of this algorithm is the Leiden algorithm. It is a community-detection algorithm for understanding the structure of complex networks. It is an upgrade to another algorithm known as the Louvain algorithm, guaranteeing the well-connectedness of communities. The Leiden algorithm has three phases, mainly: (1) local moving of nodes that creates initial partitions, (2) refining of the initial partitions, and (3) aggregating the network based on the partitions, while using the non-refined partition to create an initial partition for the aggregate network[7]. The Leiden algorithm performs constant and fast refinement until it reaches the best community partitions and makes sure that their are well-connected communities.

The Leiden algorithm is applied on a graph, making the utilization of weighted degree centrality clustering an essential initial process. The graphs that are applied with weighted degree centrality will be fed to the Leiden algorithm to create well-connected communities and relationships of the biomedical entities based on the weights and similarities.

3.4 System Design and Implementation

For this study to be useful to researchers, Phytopharmvis - a web application was developed. The system allows the users to view and analyze subnetworks based

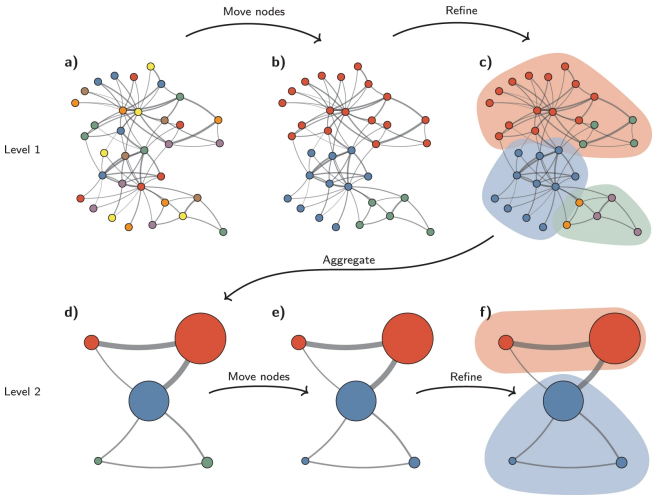

Fig. 3. A visual representation on how the Leiden algorithm works. [7]

on a plant, smd, gene or diseases of interest. The system utilizes a search field that instantiates the provision of information and visualization for that queried entity with several associations (options) to choose from.

The system uses the following online, open-source databases: a) Dr. Duke's Phytochemical and Ethnobotanical Databases for the plants and drugs datasets, b) BioGrid for the drugs and genes datasets, c) PharmKB for the drugs and genes, as well as genes and diseases datasets. All of these are in csv format and are used to create aggregated datasets for the algorithms to be used, and then uploaded to MongoDB, the online database used.

The Figure 4 shows the flow chart from the raw datasets from online resources up to the integration.

The system's front-end is developed using HTML, CSS, and Bootstrap, while its back-end is implemented using Django and JavaScript.

The graph visualization tool that is used is Sigma.JS and Graphology, which are JavaScript packages.

The online database that is used for the csv files containing the relevant information of the entities is MongoDB, which provides a total of 500MB free logical and storage memory. While the graph database that is used to store the calculated and computed graphs is Neo4j.

The querying language that is used for retrieving graphs is Cypher Query, which is specific to Neo4j and graphs, for faster graph retrieval since it is mainly focused on querying graphs.

The algorithms that are used for calculating the centrality measures and creating the clusters/communities are Degree Weighted Centrality Clustering and the Leiden Algorithm, which are both can be imported via Networkx.

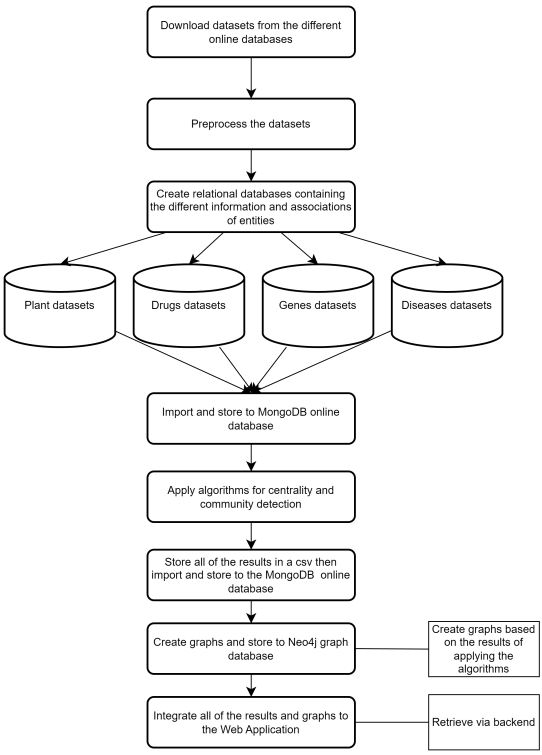

Fig. 4. Flow Chart

4 RESULTS AND DISCUSSION

4.1 Most Important Entities

Figure 5 shows the top 20 central nodes for a plant-plant network based on common smd associations; as well as the Top 20 central nodes for an smd-smd network based on common plant associations. Figure 6 shows the top 20 central nodes for an smd-smd network based on common gene associations; as well as the top 20 central nodes for a gene-gene networks based on common smd associations. Figure 7 shows the top 20 central nodes for a gene-gene network based on common disease associations; as well as the top 20 central nodes for a disease-disease network based on common gene associations.

In Figure 8, we see the top 20 central plants for a plant-gene network based on common smd associations; the top 20 central genes for a gene-plant network based on common smd associations; the top 20 central smds for an smd-disease network based on common gene associations; and the top 20 central diseases for an disease-smd network based on common gene associations.

Fig. 5. (Left)Top 20 central nodes for a plant-plant network based on common smd associations; and (Right) Top 20 central nodes for an smd-smd network based on common plant associations

Plants The following plants appear to be very central as they appear in both top 20 central plants in the plant-plant network based on common smd associations and top 20 central plants in the plant-gene network based on common smd associations : *Daucus carota, Apium graveolens, Zea mays, Citrus sinensis, Vitis vinifera, Ficus Carica, Malus domestica, Glycine max, Citrus reticulata.*

Small-Molecule Drugs The following smds appear to be very central as they appear in two of the following lists: the top 20 central smds in the smd-smd network based on common plant associations, top 20 central smds in the smd-smd network based on common gene associations, and the top 20 central smds in the smd-disease network based on common gene associations : *Olanzapine, Doxepin, Desipramine, Clozapine, Risperidone, Clomipiramine, Selective Serotonine reuptake inhibitors, Atorvastatin.*

Genes The following genes appear to be very central as they appear in two of the following lists: the top 20 central genes in the gene-gene network based on common smd associations, top 20 central genes in the gene-gene network based on common disease associations, and the top 20 central genes in the gene-plant network based on common smd associations : *ABCB1,CYP3A4, ABCC2, NR1I2, COMT, SLCO1B1, MTHFR.*

Diseases The following diseases appear to be very central as they appear in both top 20 central diseases in the disease-disease network based on common gene associations and top 20 central diseases in the disease-smd network based on

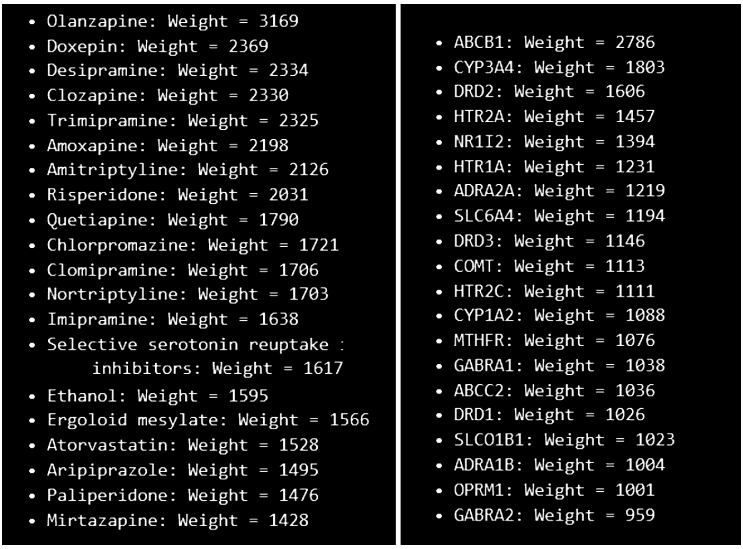

Fig. 6. (Left)Top 20 central nodes for an smd-smd network based on common gene associations; and (Right) Top 20 central nodes for a gene-gene networks based on common smd associations

common gene associations : *Breast neoplasms, Schizophrenia, Neoplasms, HIV Infection, Epilepsy, Hypertension, Colorectal neoplasms, Drug Toxicity, Pain, Kidney transplantation, Precursor Cell lymphoblastic, Depression, Ovarian neoplasms, Rheumatoid arthritis, Toxic liver disease, Major depressive disorder, Non-small-cell lung carcinoma.*

4.2 Communities

The communities are determined by implementing the Leiden algorithm onto the weighted degree centrality measured graphs. Communities composed of single elements were not included. There is a total of 8 communities in the plant to plant association via common smds, a total of 83 communities in the smds to smds association via common genes, 81 total communities in the smds to smds association via plants, a total of 11 communities in the gene to gene association via smds, 53 communities in the gene to gene association via common diseases, and a total of 10 communities in the diseases to diseases association via common genes. The distribution of the sizes of the communities are shown in Figure 9.

4.3 System Results

As was mentioned, a web-based system was developed to help researchers as they explore subbetworks of plants, smds, genes and diseases. Figure 10 shows the landing page of the system, basically its home page where it shows a brief description of the system.

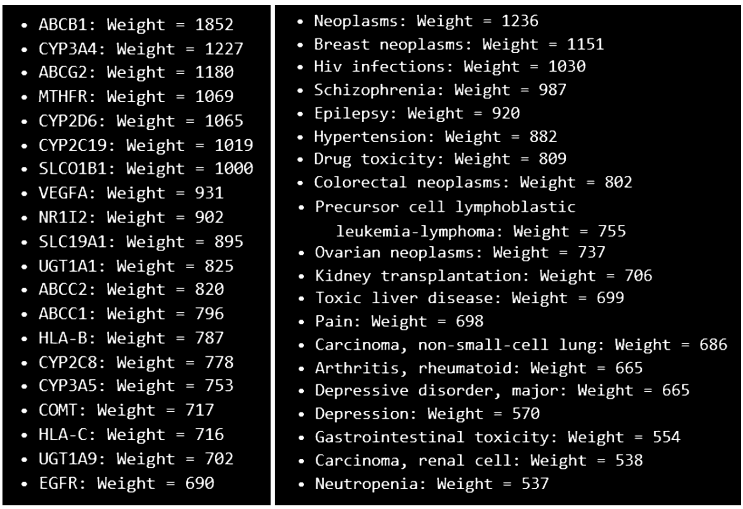

Fig. 7. (Left)Top 20 central nodes for a gene-gene network based on common disease associations; and (Right) Top 20 central nodes for a disease-disease network based on common gene associations

Figure 11 shows a sample of weighted monopartite association which shows: the retrieved edge weights for threshold value suggestions (left panel - topmost), the top central/important nodes within the association chosen (left panel - middle), sample chosen community which shows the list of entities in Community 1 of the chosen association (left panel - bottom), the graph that is retrieved after clicking "Equal" or "Greater" if there is a search input and a threshold input (middle), the various information about the searched entity (bottom panel - left and middle), and the common associated entities of that connects the two nodes/entities (bottom panel - right).

Figure 12 shows a sample of the retrieved information and graph visualization for weighted bipartite association of plants and genes. For weighted monopartite and bipartite associations, the node sizes represents its centrality measure, the higher the centrality the greater the size, while the edge sizes differ depends on the number of common associated entities between the tow nodes/entities.

Taking advantage of the latest trends in network implementation and analysis, which is the usage of graph database for faster retrieval of graphs, really provides a fast, accurate, and consistent retrieval of graphs of interest. It has still a lot of flexible features to offer that can be further used in similar studies.

Overall, the system's graph database contains a total of 338,433 nodes and 5,469,024 edges. During graph retrieval and visualization, nodes with high centrality measure appear to be larger than those with smaller centrality measure. The edges, on the other hand, also varies in thickness but is restricted to a certain value of thickness to avoid overlapping with one another if an edge has a great value. Results seen as graphic visualizations can also be observed clean

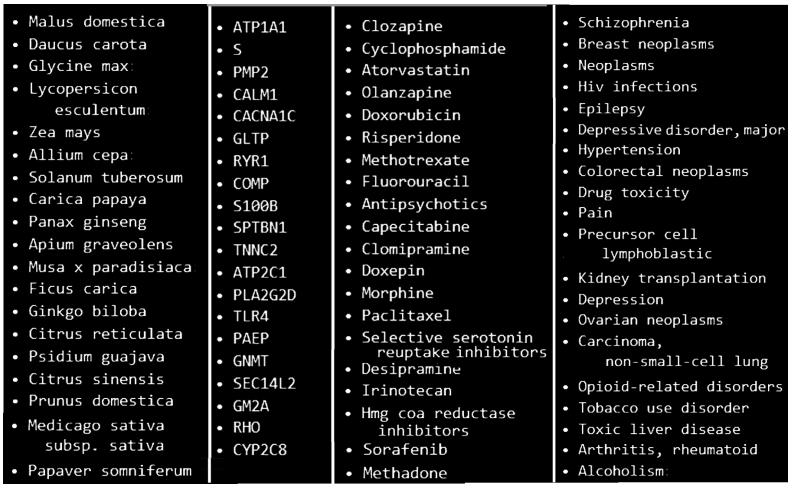

Fig. 8. (Extreme left)Top 20 central plants for a plant-gene network based on common smd associations; (Center left)Top 20 central genes for a gene-plant network based on common smd associations; (Center right) Top 20 central smds for an smd-disease network based on common gene associations; and (Extreme right) Top 20 central diseases for an disease-smd network based on common gene associations

	Number of Communities (x)				
Association Networks	$x > 1{,}000$	$1{,}000 \geq x > 100$	$100 \geq x > 10$	$10 \geq x > 1$	TOTAL
plant to plant via common smds	0	5	0	3	**8**
smd to smd via common genes	1	3	9	64	**83**
smd to smd via common plants	5	17	23	36	**81**
gene to gene via common smds	0	5	3	3	**11**
gene to gene via common diseases	0	10	7	36	**53**
disease to disease via common genes	0	5	1	4	**10**

Fig. 9. Communities determined among the association networks using the Leiden algorithm

when limited with threshold values. Unweighted bipartite associations also gives correct and clean graph visualizations.

5 CONCLUSION AND RECOMMENDATIONS

Using network science to model the relationships between plants, small-molecule drugs, genes and diseases, particularly in the identification of key entities and detection of communities is a research area that can augment and enhance the work of those engaged in phytomedical and pharmacogenomic research. It can reveal relational knowledge and insights based on actual data that may not be obvious to many researchers.

Open-source databases were used to provide data for the constructiono of the association networks, particularly, Dr. Duke's Phytochemical and Ethnobotani-

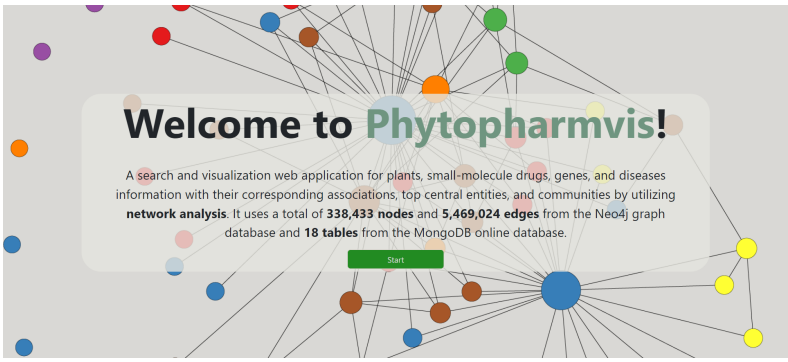

Fig. 10. Landing Page of the System

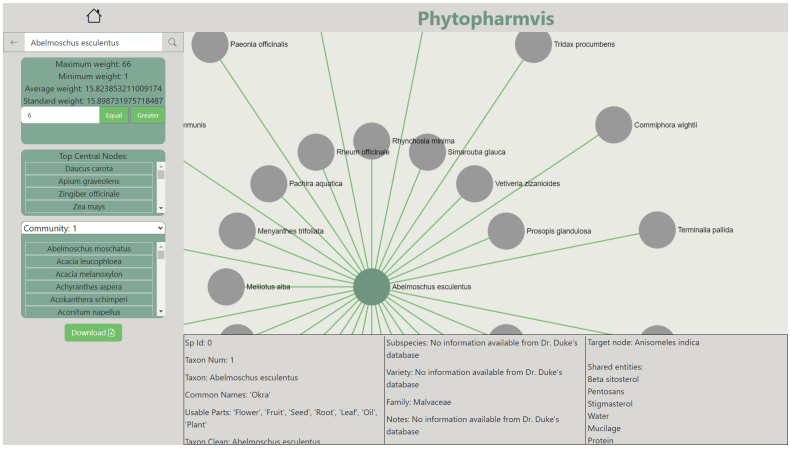

Fig. 11. Sample Retrieved and Visualized Graph with Clicked Edge Information (Edge Weights Equal to 6)

cal Databases for the plants and drugs datasets; BioGrid for the drugs and genes datasets; and PharmKB for the drugs and genes, as well as genes and diseases datasets. The networks were then analyzed using the weighted degree centrality and the Leiden algorithm for community detection. A web application was also developed, which allows the users to view and analyze subnetworks based on certain plants, smds, genes or diseases of interest. It is hoped that the results of this study as well as the tool developed will help in the field of health and medicine, specifically in phytomedicine research. It poses and associates natural sources of drugs from plants, suggests alternatives (ideas) in medicine/medicinal provisions, introduces a tool for gene and disease research, and helps in discovering or repurposing of drugs based on similarities of associations. These make this study capable for future use and further researching in this field.

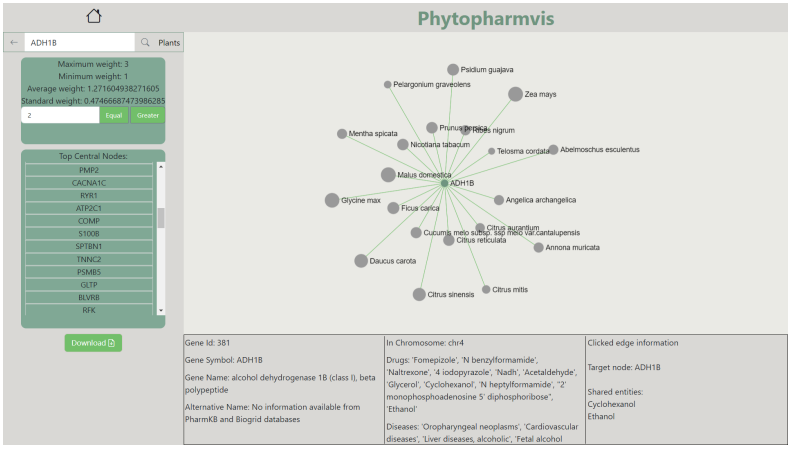

Fig. 12. Sample Retrieved Information and Graph Visualization of Weighted Bipartite Association (for Genes)

The system can be improved by implementing a feature where a different type of users to the system can add new entities with their information in the system's relational database, process them so that they create similarity edges with other entities, then convert and upload the nodes and edges to the graph database. This used can be an "admin" that will maintain and update the databases. Other data visualization libraries or packages can be tried to make an easier way of coding the interactivity features of the graph.

References

1. Dutta, S., et al: Graph-based clustering technique for microblog clustering. Science Direct (2023).
2. Elumalai, A., et al: Graph theory applications in computer science and engineering. Malaya Journal of Matematik (2020).
3. Harnoune, A., et al: BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Computer Methods and Programs in Biomedicine Update (2021).
4. Groza, V., "Drug repurposing using modularity clustering in drug-drug similarity networks based on drug–gene interactions," Pharmaceutics, 13(12), 2117, (2021).
5. Recario, J. M., et al: PlaPhy-SMD: A Database System of Plants and their Phytochemicals Integrated with Small-Molecule Drugs and their Condition and Gene Associations with Clustering and Product Ingredient Matching. IEEE Conference Publication — IEEE Xplore (2020).
6. Srivastava, A., et al: Phytomedicine. Science Direct (2019).
7. Traag, V., et al: From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports (2019).
8. Wong, M., "Search and visualization of gene-drug-disease interactions for pharmacogenomics and precision medicine research using genedive," Journal of biomedical informatics, (2021).
9. Xiong, G., "Ddinter: an online drug–drug interaction database towards improving clinical decision-making and patient safety," Nucleic acids research, (2022).