Network and Citation Visualization of Biomedical Research Publications

Ramesh Singh[1], Shashwat Aggarwal[2]
[1]Scientist-G, NIC, New Delhi, India
[2]B. E. (CS), Netaji Subhas Institute of Technology, New Delhi, India

ABSTRACT - Data visualization techniques proffer efficient means to organize and present data in graphically appealing forms which not only speeds up the process of decision making and pattern recognition but also enables decision makers to fully understand data insights and make informed decisions. Over time, with the rise in technological and computational resources, there has been an exponential increase in world's scientific knowledge. However, most of it lacks structure and cannot be easily categorized and imported into regular databases. This type of data is often termed as Dark Data. Data visualization techniques provide a promising solution to explore such data by allowing quick comprehension of information, discovery of emerging trends, identification of relationships and patterns etc. In this empirical research and study, we use the rich corpus of the PubMed comprising of more than 28 million citations for biomedical literature to explore and analyze lexical and textual biomedical dark data using Network and Citation visualization techniques. We use VOSviewer to construct bibliometric networks for studying relationships between different entities like scientific documents and journals, researchers, and, keywords and terms. We discuss some of the techniques and methodology used by VOSviewer to preprocess enormous datasets and construct large scalable networks efficiently. The paper concludes with a discussion of the limitations and future applications of network and citation visualizations.
KEYWORDS - Network Data, Bibliometric, Bibliographic Coupling, PubMed, Co-Citation, VOSviewer, Dark Data.

I. INTRODUCTION

In today's data centralized world, the practice of data visualization has become an indispensable tool in numerous domains such as *Research*, *Marketing*, *Journalism*, *Biology* etc. Data visualization is the art of efficiently organizing and presenting data in a graphically appealing format. It speeds up the process of decision making and pattern recognition, thereby enabling decision makers to make informed decisions.

With the rise in technology, the data has been exploding exponentially, and the world's scientific knowledge is accessible with ease. There is an enormous amount of data available in the form of scientific articles, government reports, natural language, and images that in total contributes to around 80% of overall data generated as shown in an excerpt from The Digital Universe (IDC, The Digital Universe, 2012) in Figure 1. However, most of the data lack structure and cannot be easily categorized and imported into regular databases. This type of data is often termed as Dark Data. Data visualization techniques proffer a potential solution to overcome the problem of handling and analyzing overwhelming amounts of such information. It enables the decision maker to look at data differently and more imaginatively. It promotes creative data exploration by allowing quick comprehension of information, the discovery of emerging trends, identification of relationships and patterns etc.

Over time, data visualization techniques have been used in a great number of domains, but the domain that has received major attention recently is text. Professional users like scholars, research institutions, and funding agencies have become more and more interested in the textual domain. There are numerous techniques used for visualization of text as summarized in Figure 2. All these techniques can be categorized into various subdomains depending on their use case, for instance, geometric, clustering based, graph-based etc. One such technique that has been used extensively in the past, especially in the analysis of research documents and case studies is Network and Citation visualization.

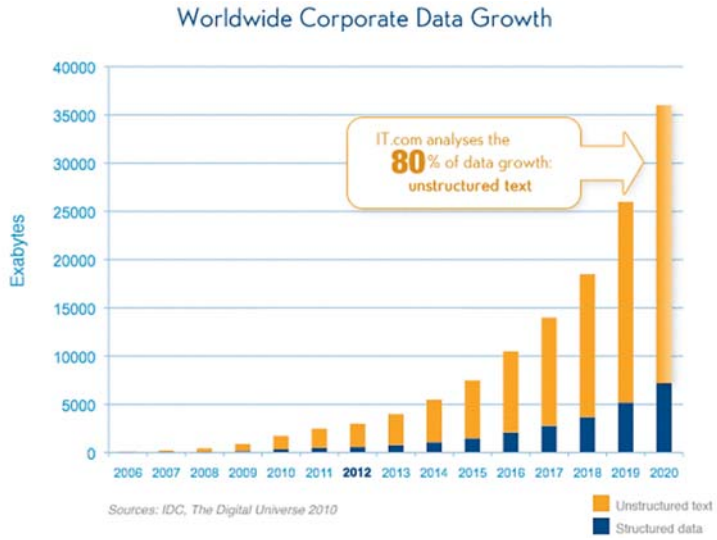


Figure 1. Worldwide growth of corporate data categorized by structured and unstructured/dark data over the past decade.

Network visualization also called as Network graphs are often used to visualize complex and convoluted relations between an enormous amount of entities. They represent information in a hierarchically structured manner through an interconnected network of entities highlighting the correlation between them. At its most basic level, a network graph consists of nodes and edges. Nodes represent the entities, and edges represent the relationship between those entities. Edges in the graph can be directed or undirected. Directed edges indicate the flow of information from one node to another. Undirected edges, on the other hand, indicate the presence of a bidirectional relationship between the two nodes.

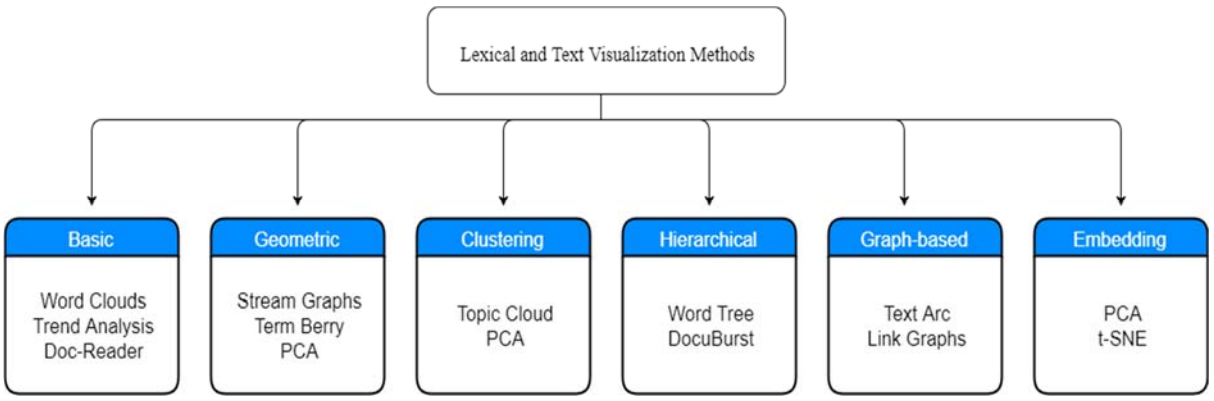


Figure 2. Hierarchy of Lexical and Text Visualization Techniques used.

One of its variants, Citation networks is used extensively in the field of bibliometrics. “Bibliometrics is the scientific analysis of publications that seek to identify the major fields of study within and across scientific disciplines and the most influential publications, research institutions, and researchers in each of these fields.” (Belter, 2012). Citation networks proffer quick summarization and visualization of the structure inherent to a set of publications. The resulting visualizations provide insights not only into the present state of scientific research but in identifying potential future research directions and collaboration opportunities. Bibliometric or citation networks can be classified into direct citation networks, co-citation networks, and bibliographic coupling relations. Direct citation networks, also known as cross citation networks represent research documents citing each other directly as nodes in the network. These networks only offer a direct indication of relatedness between the entities. These are usually very sparse networks and hence relatively uncommon in research settings.

The second variant of citation networks i.e., co-citation networks represents co-cited research documents (i.e., a pair of documents that are cited by some other group of common documents) as network entities. The greater the number of research documents citing the two documents, the stronger is the relatedness between them. White and McCain (1998) used these co-citation networks to study the researchers in the field of information science. Lastly, in bibliographic coupling, two documents are said to be coupled if both cite a common research document (Kessler, 1963). In other words, the more common the references two documents have, the stronger is the coupling relation between them. Bibliographic coupling has continued to receive increasing attention over time (eg. Boyack and Klavans, 2010; Jarneving, 2007).

Network and citation visualization techniques have been used extensively in research settings to analyze overwhelming sets of information. For instance, Ke et al. (2004) used “citation analysis to identify major research topics, co-authorships, and trends in the InfoVis Contest dataset that contains 614 papers published between 1974 and 2004.” They created both static and dynamic visualizations using the open source network and citation viewer, Pajek (Batagelj and Mrvar, 1998). Perianes-Rodríguez et al. (2010) performed “structural analysis to generate co-authorship networks and identify research groups based on factorial analysis of the raw data matrix and similarities in the choice of co-authors.” Further, Bauer-Mehren et al. (2010) discussed about a Cytoscape plugin (Shannon et. al, 2003) called DisGeNet used to query and analyze human gene-disease networks. “It represents gene-disease associations in terms of a bipartite graph and helps in providing gene-centric and disease-centric views of the data.” (Bauer-Mehren et al., 2010). The plugin proves to be an immense help to researcher present in the field concerning human gene-disease interactions in exploring and interpreting the genetic basis of human diseases. Shibata et al., (2011) detected “an emerging research front in a vast number of academic papers related to regenerative medicine by dividing the citation networks into clusters using the topological clustering method and tracking the positions of papers in each cluster to detect contemporary trends.” More recently, Nakazawa et al. (2015) proposed “a visualization technique for citation networks by applying topic-based paper clustering using LDA (Latent Dirichlet Allocation) to construct clustered networks of the research papers.”

In this paper, we first provide an overview of the dataset that we used in our study, i.e. PubMed. We then survey some of the most popular tools available to construct the bibliometric networks for studying relationships between different entities like scientific documents and journals, researchers, and, keywords and terms. We then specifically focus on the VOSviewer tool, discussing the techniques and methodology used by the tool to construct large scalable networks efficiently and finally generate and analyze networks for our dataset. The paper concludes with a discussion of the limitations and future applications of network and citation visualizations.

II. OVERVIEW OF PUBMED DATASET

“PubMed comprises of more than 28 million citations and abstracts for biomedical literature from MEDLINE, life science journals, and online books.” (https://www.ncbi.nlm.nih.gov/pubmed/, 2018). It is an open source database developed and maintained by the National Center for Biotechnology Information (NCBI). In addition to free access to MEDLINE, PubMed also provides links to free full-text articles provided by PubMed Central and third-party websites, advanced search capabilities, clinical queries search filters, special query pages and other related resources. “PubMed is a key information resource in biological sciences and medicine primarily because of its wide diversity and manual curation. It comprises of an order of three billion bases of human genome, rich meta-information (e.g. MeSH terms), detailed affiliation, etc., summing up to a total of 70GB database.” (Roberts, 2001).

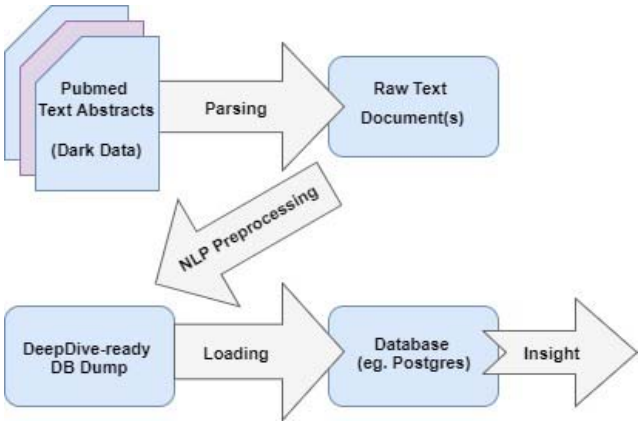


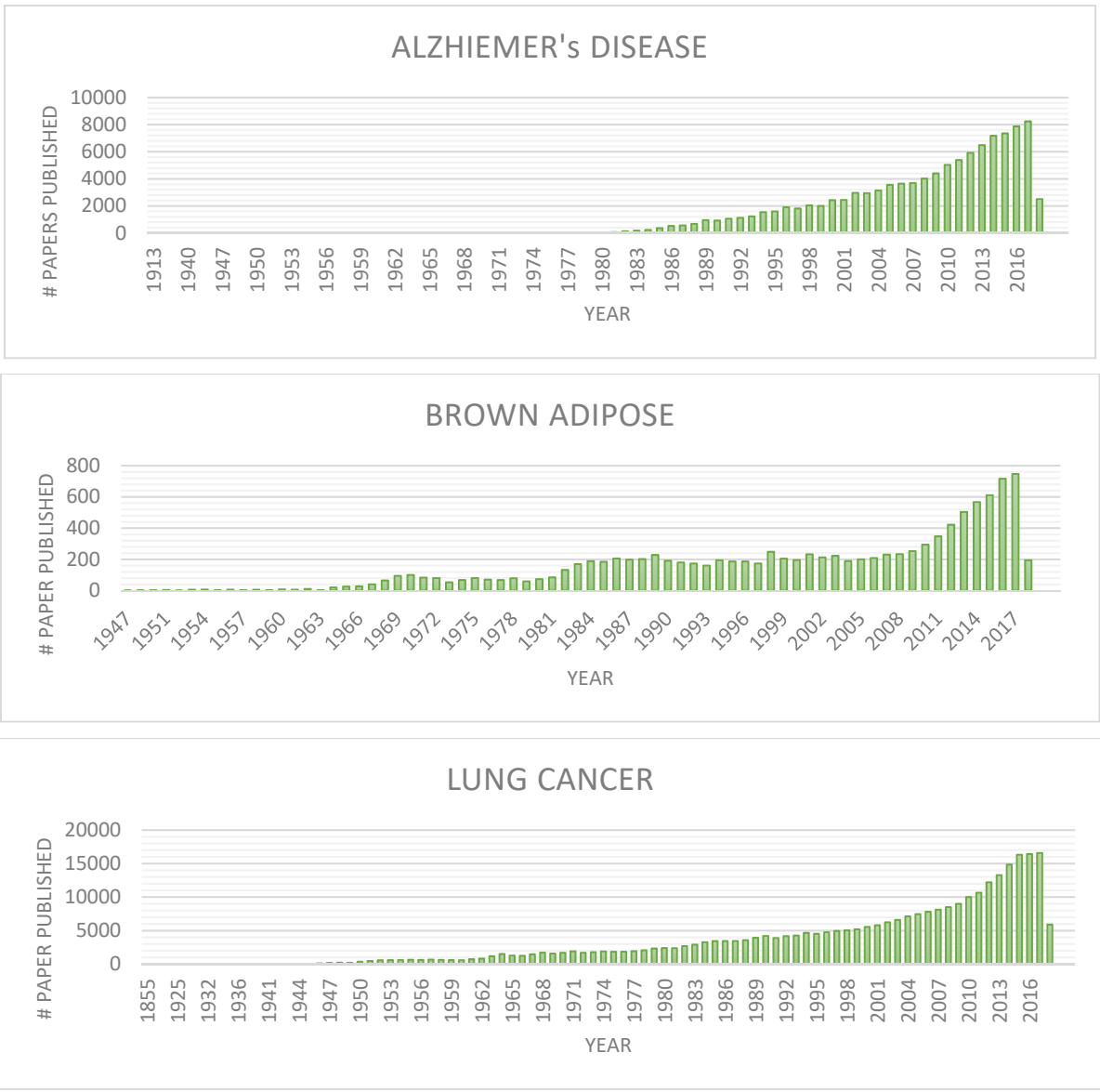
Figure 3. Pre-processing pipeline adopted by Deep Dive to convert PubMed dataset into a structured format.

As of 1 April 2018, PubMed has more than 28 million records from 5500 journals, dating back to 1966, with the earliest publication available from the year 1809. “13.1 million of PubMed's records are listed with their abstracts, and 14.2 million articles have links to full-text with around 500,000 new records being added each year. Around 12% of the records in PubMed correspond to cancer-related entries, which have grown from 6% in the 1950s to 16% in 2016. Other significant proportions of records correspond to “Chemistry” (8.69%), “Therapy” (8.39%) and "Infection" (5%).” (PubMed, Wikipedia, 2018). PubMed provides efficient methods to search the database by using author names, journal names, keywords or phrases, MeSH terms or any combination of these. It also enables users to download the fetched citations and abstracts for queried terms in various formats such as plain text form (both Summary and Abstract), XML form, PMID form, CSV form and MEDLINE form. The results are sorted according to one of the followings: publication date, author name, title of paper or journal name.

Table 1: Fundamental Statistics of PubMed Dataset as on March 2014; (http://deepdive.stanford.edu/opendata/#pmc-oa-pubmed-central-open-access-subset, 2014)

Size	70 GB	**# Sentences**	110 Million
# Documents	359,324	**# Distinct Entities**	412,593,720
# Words	2.7 Billion	**# Distinct Subjects**	412,593,720
# Distinct Literals	1,842,783,647	**# Distinct Objects**	436,101,294

Deep Dive (Deep Dive, 2017) is an open source dark data analytical platform developed by Christopher Ré and his group at Stanford which hosts pre-processed copies of several open source text databases (e.g. PubMed). The processing pipeline used by Deep Dive to create a structured database from the research documents present in PubMed is shown in Figure 3. The pipeline consists of several stages, namely: a) scraping of data in HTML/XML format (Parsing), b) striping it into plain text, c) applying basic NLP pre-processing such as tokenization, stemming, POS tagging etc to clean up the data, and d) loading the Deep Dive ready DB dump in a structured database like Postgres. The fundamental statistics of PubMed database provided by Deep Dive are reported in Table 1. These statistics help in providing an abstract overview of the database like for example the frequency count at various levels (literals, words, sentences, documents).



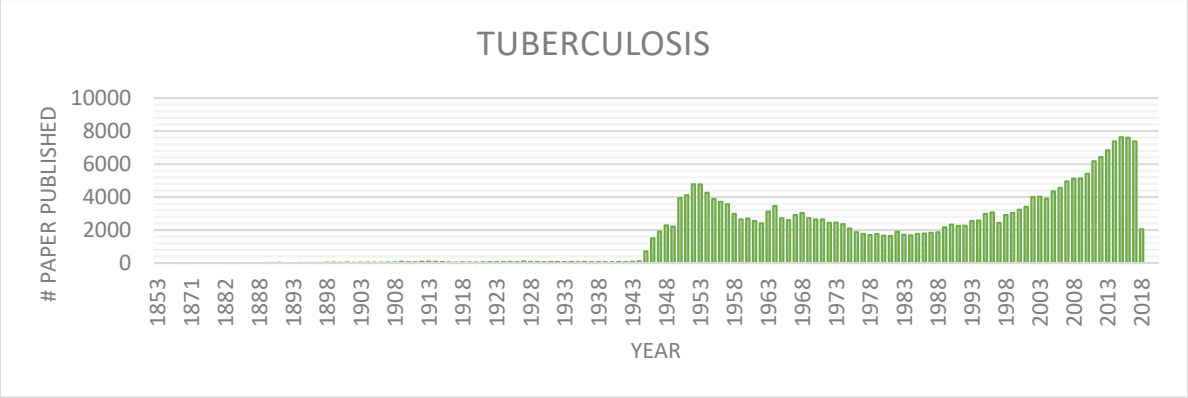


Figure 4. Timeline of PubMed Articles for queried topics: Alzheimer's Disease, Brown Adipose, Lung Cancer and Tuberculosis respectively.

In order to create network and citation visualizations on PubMed abstracts and thereafter infer and study the insights obtained from them we query the PubMed database on four broad topics namely, Alzheimer's Disease, Brown Adipose, Lung Cancer, and Tuberculosis respectively. The timelines for the number of articles published in PubMed concerning with these topics over the past century are illustrated in Figure 4. We obtain the resulting abstracts for each topic from the PubMed site in MEDLINE format. “The MEDLINE format is a tagged field format that displays all fields of a MEDLINE record.” (NIH, U.S. National Library of Medicine, 2018). As an example, consider one of the fetched abstract related with queried topic Brown Adipose, shown in Figure 5 respectively. Each field is tagged by its identifier and can be accessed by accessing the corresponding identifier. The MEDLINE format because of its organized tagged field form allows efficient and very fast processing of data. Before moving on to the generation of network and citation visualizations, we briefly survey some of the popular tools available to create network and citation visualizations in the next section.

```
PMID- 29634313
OWN - NLM
STAT- Publisher
LR  - 20180410
IS  - 1522-1555 (Electronic)
IS  - 0193-1849 (Linking)
DP  - 2018 Apr 10
TI  - NAD(+)-dependent deacetylase SIRT3 in adipocytes is dispensable for maintaining
      normal adipose tissue mitochondrial function and whole-body metabolism.
LID - 10.1152/ajpendo.00057.2018 [doi]
AB  - Mitochondrial dysfunction in adipose tissue is involved in the pathophysiology of
      obesity-induced systemic metabolic complications, such as type 2 diabetes,
      insulin resistance, and dyslipidemia. However, the mechanisms responsible for
      obesity-induced adipose tissue mitochondrial dysfunction are not clear. The aim
      of present study was to test the hypothesis that nicotinamide adenine
      dinucleotide (NAD(+))-dependent deacetylase SIRT3 in adipocytes plays a critical
      role in adipose tissue mitochondrial biology and obesity. We first measured
      adipose tissue SIRT3 expression in obese and lean mice. Next, adipocyte-specific
      mitochondrial Sirt3 knockout (AMiSKO) mice were generated and metabolically
      characterized. We evaluated glucose and lipid metabolism in adult mice fed either
      a regular-chow diet or high-fat diet (HFD), and in aged mice. We also determined
      the effects of Sirt3 deletion on adipose tissue metabolism and mitochondrial
      biology. Supporting our hypothesis, obese mice had decreased SIRT3 gene and
      protein expression in adipose tissue. However, despite successful knockout of
      SIRT3, AMiSKO mice had normal glucose and lipid metabolism and did not change
      metabolic responses to HFD-feeding and aging. In addition, loss of SIRT3 had no
      major impact on putative SIRT3 targets, key metabolic pathways and mitochondrial
      function in white and brown adipose tissue. Collectively, these findings suggest
      that adipocyte SIRT3 is dispensable for maintaining normal adipose tissue
      mitochondrial function and whole-body metabolism. Contrary to our hypothesis,
      loss of SIRT3 function in adipocytes is unlikely to contribute to the
      pathophysiology of obesity-induced metabolic complications.
FAU - Porter, Lane C
AU  - Porter LC
AD  - Washington University School of Medicine.
FAU - Franczyk, Michael P
AU  - Franczyk MP
AD  - Washington University School of Medicine.
FAU - Pietka, Terri
AU  - Pietka T
AD  - Washington University School of Medicine.
FAU - Yamaguchi, Shintaro
AU  - Yamaguchi S
AD  - Department of Internal Medicine, Keio University School of Medicine.
FAU - Lin, Jonathan B
```

Figure 5. Example MEDLINE Format fetched article from PubMed for query search *“Brown Adipose”*.

III. A SURVEY OF NETWORK AND CITATION VIEWERS

In this section, we take a brief survey of some of the open source and free software tools that are available for dark data analytics primarily focused on network and citation visualizations. Some of these tools are for general network analysis while some are specifically designed for the purpose of bibliometric analysis. For a more comprehensive overview of network tools, we suggest the reader to refer Cobo et al. (2011). We first survey the general network analysis tools, Pajek and Gephi. We then overview the tools specifically focusing on bibliometric analysis.

Pajek (Batagelj and Mrvar, 1998) is one of the earliest network analysis tool used primarily for the analysis and visualization of graphs and networks with a large number of vertices. It provides a number of techniques like clustering, primary path analysis, graph-based visualizations etc. It also provides the capability to export visualizations to several different viewers such as Gephi, VOSviewer etc. Gephi (Bastian et. al, 2009) on the other hand, is an open source exploration software which focuses more on network visualization aspect rather than the network analysis part. It is capable of handling and working with very large datasets efficiently to produce powerful visualizations. Gephi has a support for a number of different applications such as exploratory data analysis, link analysis, social network analysis, biological network analysis, poster creation etc. In addition to high-performance engine and support for several use cases, it also provides with several metrics such as centrality (used in sociology to indicate how well a node is connected in a graph), density, path length, diameter, HITS, clustering coefficient etc.

Apart from Pajek and Gephi, there are certain tools that specifically focus on the bibliometric visualization and analysis aspect. Here we look at three of them namely, Cytoscape (http://www.cytoscape.org), CitNet Explorer (Van Eck and Waltman, 2014[a]) and VOSviewer (Van Eck and Waltman, 2010). Cytoscape is an open source software visualization platform which was initially designed for visualizing molecular interaction networks and biological pathways, but at present can be used as a general platform for complex network analysis and visualization. It provides a basic set of features for data integration, analysis, and visualization with an additional set of features available as Apps that can be freely availed from Cytoscape App Store.

CitNet is another software tool developed by Leiden University for visualizing and analyzing citation networks of scientific publications. The tool provides support for directly importing information from the Web of Science database and visualizing the same as a citation network. In addition to that, it provides support for several applications such as analysis of the development of a research field over time, identifying the literature on a research topic, exploring the publication oeuvre of a researcher, supporting literature reviewing and many more.

Lastly, VOSviewer is "an open source software tool used for constructing and visualizing bibliometric networks. These networks may include journals, scientific documents, or published articles etc. Visualizations can be constructed based on citation, bibliographic coupling, co-citation, or co-authorship relations." (Van Eck and Waltman, 2010). In addition to these functionalities, VOSviewer also supports text mining capability that to visualize co-occurrence networks of important terms extracted from a body of scientific literature. In the next section, we comprehensively explore the inner working of VOSviewer and utilize it for our use case of extracting useful insights from PubMed documents and creating network and citation visualizations to represent those insights systematically.

IV. OVERVIEW OF VOSVIEWER CITATION AND NETWORK VIEWER

IV.I. CREATING MAPS BASED ON NETWORK DATA

"VOSviewer is used to construct networks of scientific publications, scientific journals, researchers, research organizations, countries, keywords, or terms." (Van Eck and Waltman, 2017). Each map consists of a network of objects of interests, also known as entities. Entities in these maps such as research documents, authors, or keywords are connected by citation (co-citation and bibliographic coupling), co-authorship, or co-occurrence links. Each link has a strength associated with it, represented by a positive numerical value. The higher this value, the stronger is the link.

Further, each entity is grouped into a non-overlapping and exhaustive cluster. Entities have various attributes associated with them for instance, the weight attribute of the entity. The weight of an entity indicates the importance of that entity in the network. An entity with a higher weight is regarded as more important than an entity with a lower weight and hence shown more prominently in the visualization. VOSviewer supports three types of visualizations namely, the network visualization, the overlay visualization, and the density visualization. In the next section, we briefly discuss about these visualizations.

IV.II. NETWORK OVERLAY AND DENSITY VISUALIZATION

In network visualization, each item is represented as a circular node along with its label. Every node is interconnected with other nodes through edges representing links between those items. The size of the node is determined by the weight of the entity. The higher the weight, the larger the node size. The color of the node is determined by the cluster to which the entity belongs. The distance between any two nodes in the visualization indicates the strength of the relationship between those nodes. The closer the nodes are to each other, the stronger they are correlated. The overlay visualization is like the network visualization except that nodes are colored depending on a criterion. The criterion can be the scores of an entity or a user-defined color.

In density visualization, there are two variants namely, the item density visualization and the cluster density visualization. In the item density visualization, entities are represented by their label in a similar way as in the network visualization. Each point in the item density visualization has a color that indicates the density of entities at that point. By default, colors range from blue to green to red. The larger the number of items in the neighborhood of a point and the higher the weights of the neighboring items, the closer the color of the point is to red. The cluster density visualization is similar to the item density visualization except that the density of items is displayed separately for each cluster of items. In the cluster density visualization, the color of a point in the visualization is obtained by mixing the color of different clusters. The weight given to the color of a certain cluster is determined by the number of items belonging to that cluster in the neighborhood of the point.

IV.III. PIPELINE OF VOSVIEWER

THESAURUS FILES

A VOSviewer thesaurus file is used to perform data cleaning to create maps based on bibliographic or text data. The thesaurus file performs various functions such as merging different variants of author names, journal and organization names, synonyms, terms, spelling differences (e.g., '*color'* and *'colour'*) and cited references etc. It is useful in several circumstances like when the names of researchers are written or when the references are cited in different formats in different documents. VOSviewer uses the thesaurus file to identify these outliers like different names, in fact, refer to the same researcher and perform data cleaning.

IDENTIFICATION OF TERMS

The next step after data cleaning is the identification and selection of terms to be included in the network. Several natural language processing (NLP) based algorithms are utilized by VOSviewer for identification and selection of terms as discussed below.

1. ***Removal of copyright statements***: Text data provided to VOSviewer usually consists of titles and abstracts of scientific publications. Abstracts that contain copyright statements are identified and the copyright statements are removed.
2. ***Sentence detection***: Text data is tokenized into sentences using a NLP based sentence detection algorithm VOSviewer uses the sentence detection algorithm provided by the Apache OpenNLP library (https://opennlp.apache.org/).
3. ***Part-of-speech tagging***: After sentence tokenization, part-of-speech tagging is performed on individual sentences. VOSviewer use the part-of-speech tagging algorithm provided by the Apache OpenNLP library. Using this algorithm, each word is assigned a part of speech tag, such as verb, noun, adjective, preposition etc.
4. ***Noun phrase identification***: The next stage in the pipeline is the process of noun phrase identification. VOSviewer defines a noun phrase as "a sequence of one or more consecutive words within a sentence such that the last word in the sequence is a noun and each of the other words is either a noun or an adjective." (Van Eck and Waltman, 2017). To identify noun phrases, VOSviewer considers only the longest noun phrases that can be found in a sentence. Shorter noun phrases embedded within longer ones are ignored.
5. ***Noun phrase unification***: Lastly, the unification of noun phrases is done by removing non-alphanumeric characters, normalizing case to lower case, and by converting plural noun phrases to singular. Plural to singular conversion is done by examining the last word in a noun phrase and making conversion accordingly.

SELECTION OF TERMS

Finally, after the term identification stage which provides a set of noun phrases or terms, the next stage does the selection of terms. The selection of terms is done by excluding terms having fewer occurrences with a low relevance score (e.g. terms like 'introduction', 'results', 'conclusions' etc.), and excluding some terms manually. The selected terms thus are used to create the network visualization.

V. TOPICS FOR CITATION VISUALIZATION

In this section, we use VOSviewer to create network and citation visualization on the documents we queried from PubMed in MEDLINE format. We create two types of plots, i.e., Co-Authorship and Co-occurrence Word Network plots. In addition to these, we also create Item and Cluster Density plots from the corresponding Word Network plots.

For Co-Authorship networks, we choose the parameter of *full counting*, i.e. each link contributes equally and *Authors* as the unit of analysis. We select the minimum number of documents of an author to be 10 as the threshold to limit our network to mentions of authors who have contributed at least 10 documents related to the topic for which network map is created. Finally, from the authors shortlisted we select top 500 authors to be visualized in our network based upon their total link strength which indicates the total strength of the co-authorship links of a given researcher with other researchers.

For the Co-occurrence based Word Networks and Item and Cluster Density visualization plots, we first select the option to ignore copyright statement to get rid of unwarranted text, we extract text from both title and abstract fields of the MEDLINE document and we select the option of *full counting*, to count all the occurrences of a term in a document. We filter the less significant terms from more significant terms by setting the minimum number of occurrence of a term barrier to 8. For each of the filtered term, VOSviewer calculates a relevance score, which represents the specificity of a term towards the topics covered by the text data. We select the top 80% most relevant terms depending upon the relevance score metric to be displayed in our network. We select minimum cluster size to be 2 and Association strength to be the normalization method for the layout algorithm for both visualizations discussed above.

V.I. ALZHEIMER'S DISEASE

Figure 6 shows the Co-Authorship network for PubMed documents related to the topic *Alzheimer's Disease*. From the figure, we can observe that some authors like *Bennett Da*, *Blennow K*, *Perry G, Zhang Y*, *Wang Y* have bigger node sizes as compared to other authors thereby indicating a greater proportion of work contributed by these authors in the queried field of work. From the same figure we can also observe the potential clusters of authors depending upon the papers they have co-authored and the topics their study focusses upon. Authors in the purple and yellow cluster usually work with other authors of their clusters only while the work of authors of red, blue and cyan clusters are uniformly interspersed between different clusters.

From the distance between two authors in the visualization and the thickness of links connecting them, the relatedness of the authors can be inferred. In general, the closer two authors are located to each other or thicker the link connecting them is, the stronger their relatedness.

Various useful insights can be gathered from the co-occurrence word network shown in Figure 7. Firstly, we can infer that Alzheimer's disease is related to the brain due to presence cooccurring terms such as *'brain', 'memory', 'cognition'* etc. We can also infer that males are more vulnerable to Alzheimer disease as compared to females. Potential age group suffering from Alzheimer disease can also be identified as the middle age to old age group. Various side effects of Alzheimer disease can also be found such as *memory disorders, dementia, depression etc.*

Finally, from the item and cluster density plots shown in Figure 8 and 9, density of terms with respect to semantically related cooccurring terms present in their neighborhood and the density of terms for the potential clusters can be observed.

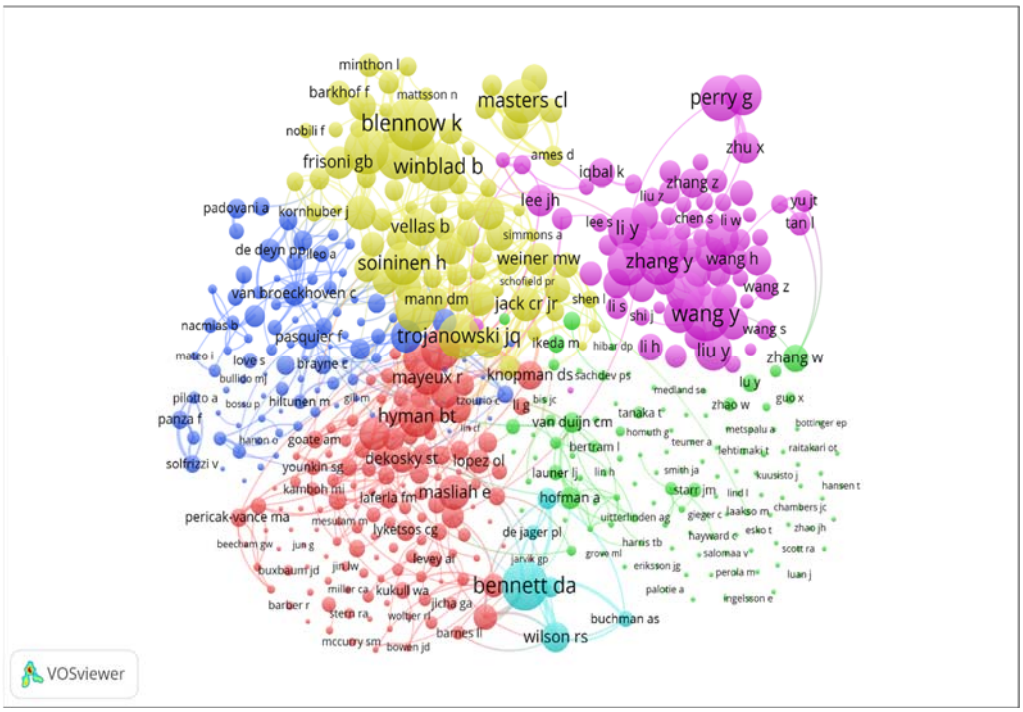


Figure 6. Co-Authorship Network Visualization of documents related to topic 'Alzheimer's Disease' queried from PubMed.

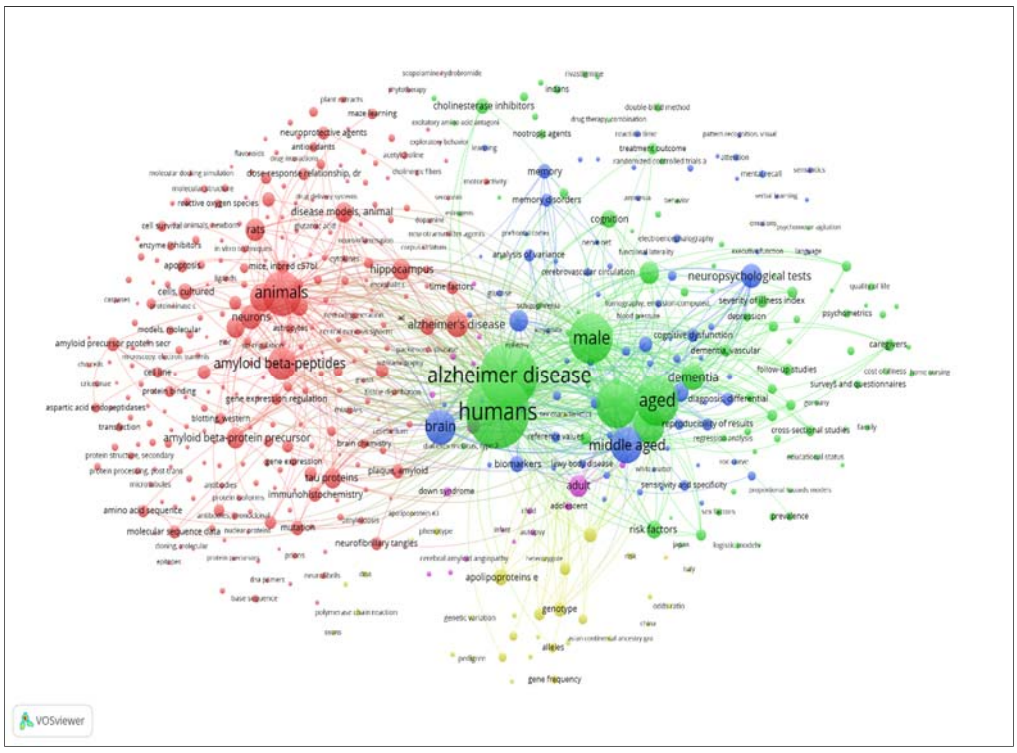


Figure 7. Co-Occurrence Word Network Visualization of documents related to topic 'Alzheimer's Disease' queried from PubMed.

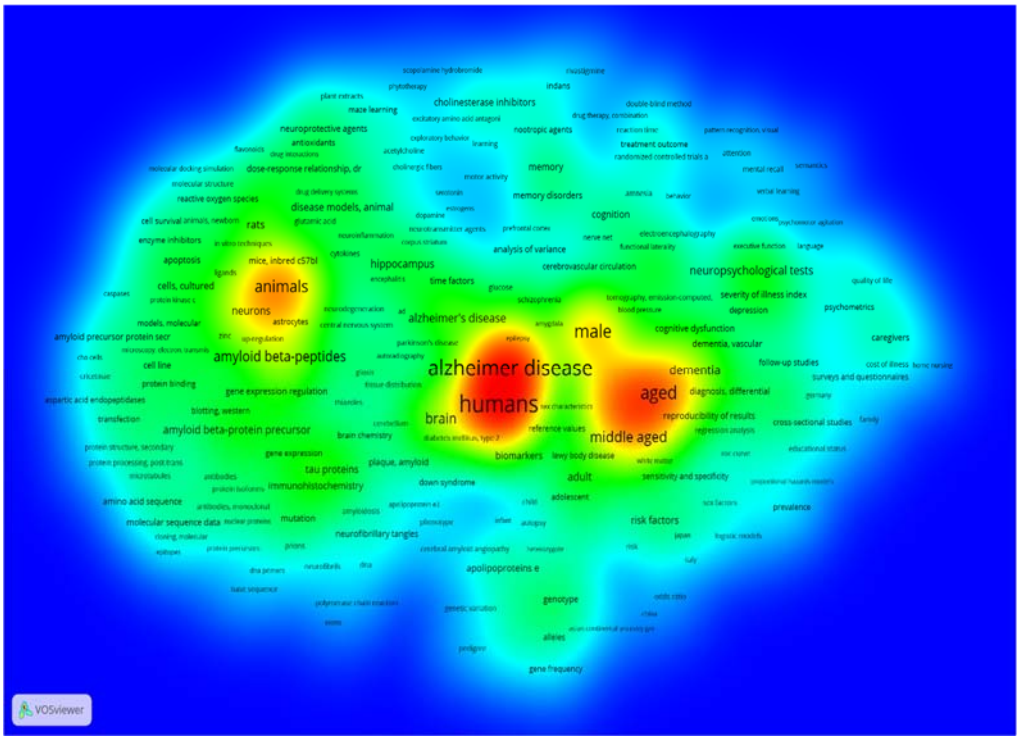


Figure 8. Item Density Visualization of documents related to topic 'Alzheimer's Disease' queried from PubMed.

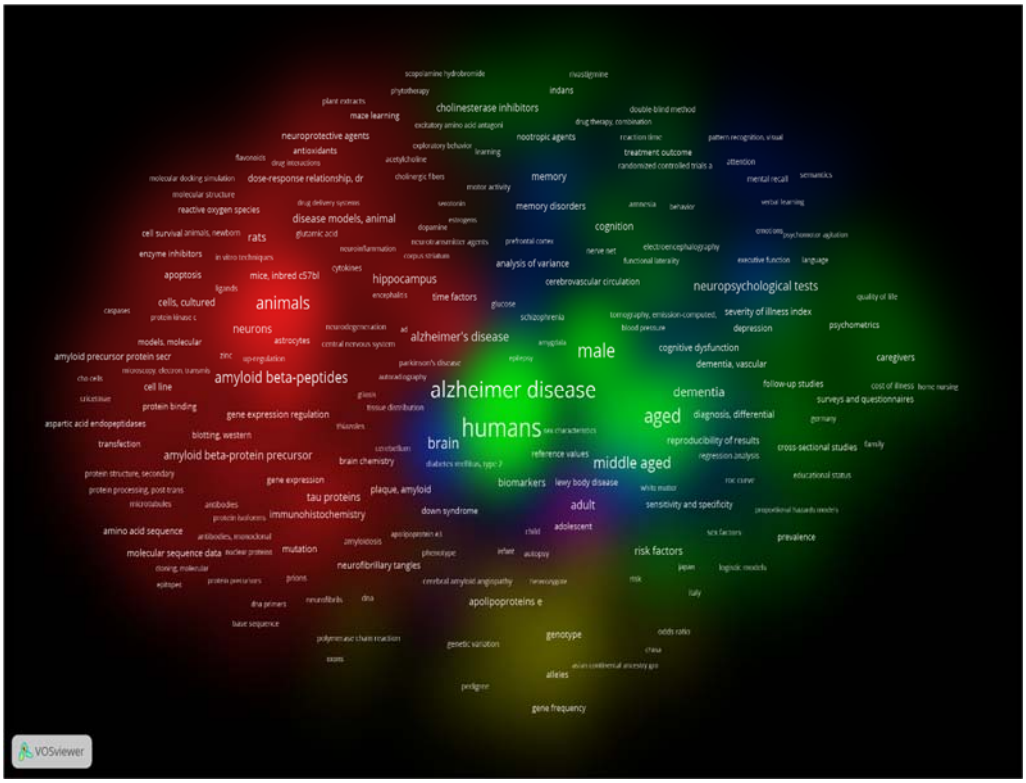


Figure 9. Cluster Density Visualization of documents related to topic 'Alzheimer's Disease' queried from PubMed.

V.II. BROWN ADIPOSE

Figure 10 shows the Co-Authorship network for PubMed documents related to the topic *Brown Adipose*. In this case, we can observe that mostly all the authors have similar node sizes with an exception of one or two authors such as *Nedergaard J* and *Saito M* which are densely connected with their neighboring co-authors. The Co-Authorship network for brown adipose suggests a steady and uniform amount of work done by the researchers in the field over the years.

The word network for brown adipose shown in Figure 11 provides certain useful insights such as brown adipose is correlated with body tissues and cells. It is more significant in animals as compared to humans. In humans, males are more vulnerable to it as compared to females. Some of its side effects and causes can also be studied from the graph such as obesity, blood pressure, age factors, sex factors etc. One particular term that is prominent is cold temperature and body temperature regulation indicating a correlation between body temperature and brown adipose.

Finally, from the item and cluster density plots shown in Figure 12 and 13, the density of terms with respect to semantically related cooccurring terms present in their neighborhood and the density of terms for the potential clusters can be observed.

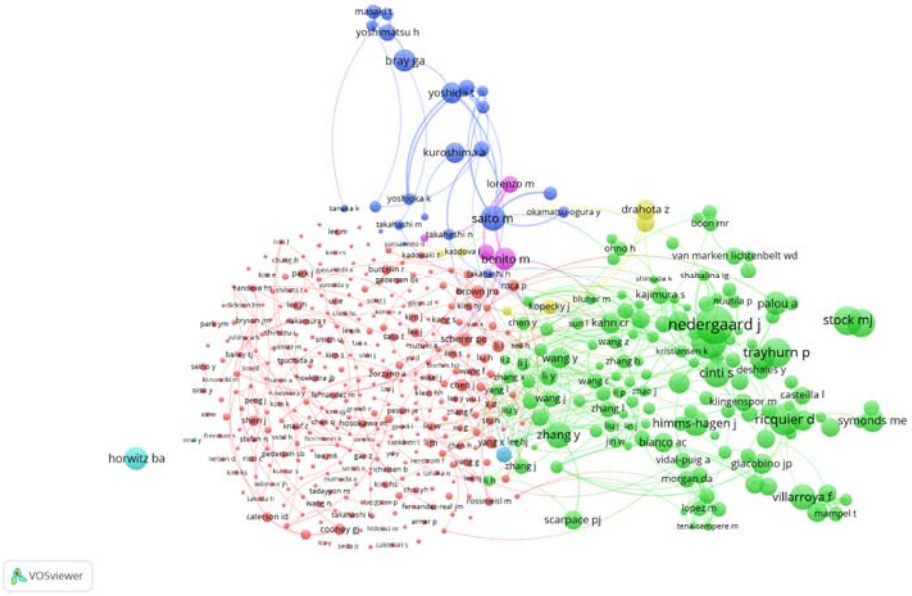


Figure 10. Co-Authorship Network Visualization of documents related to topic 'Brown Adipose' queried from PubMed.

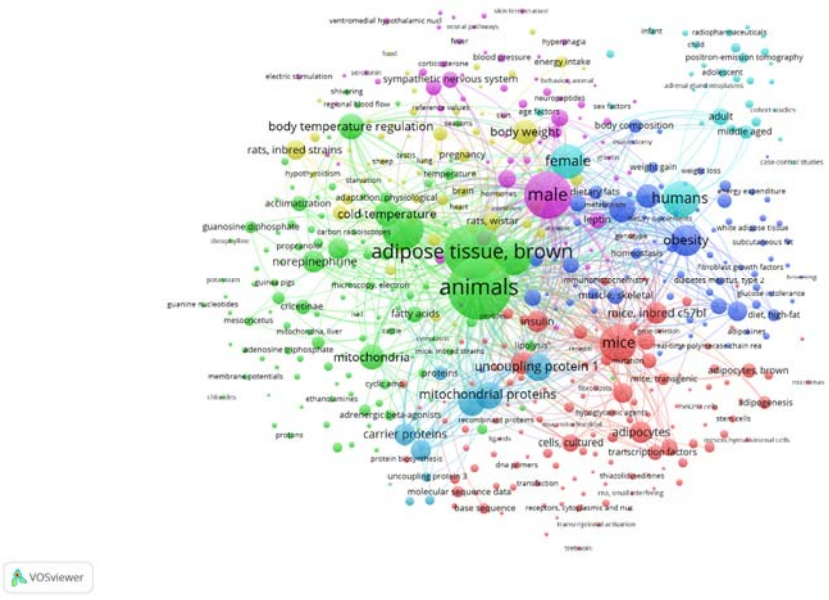


Figure 11. Co-Occurrence Word Network Visualization of documents related to topic 'Brown Adipose' queried from PubMed.

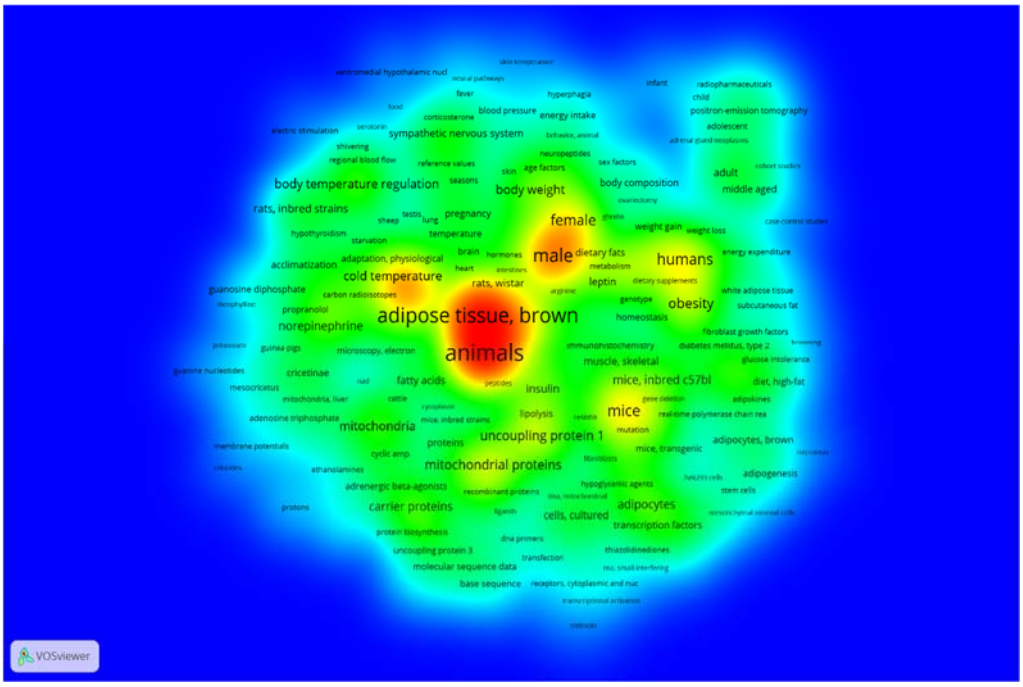


Figure 12. Item Density Visualization of documents related to topic 'Brown Adipose' queried from PubMed.

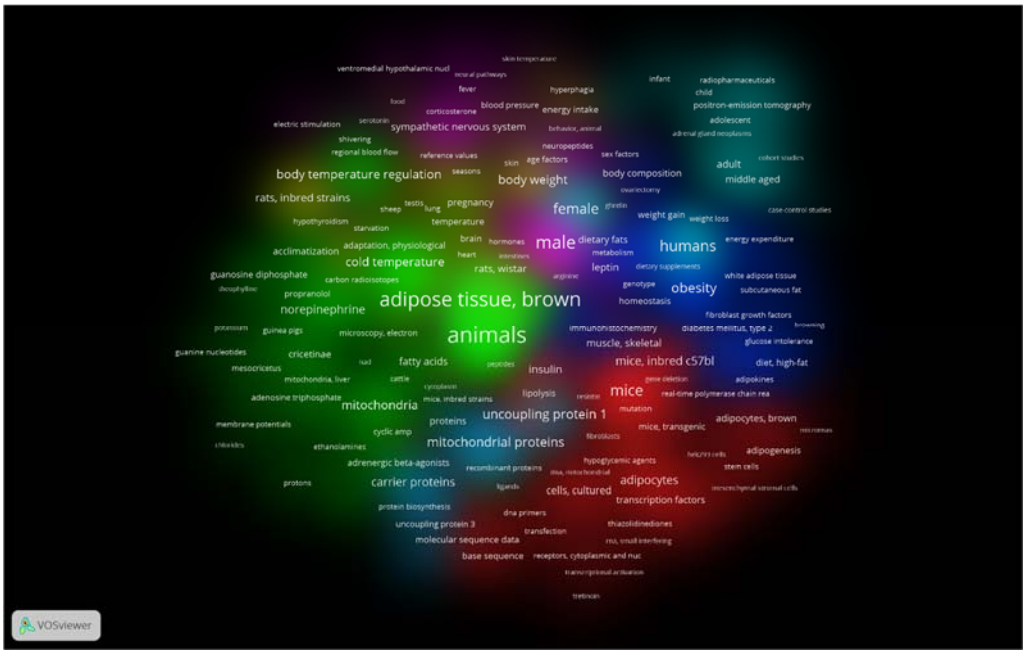


Figure 13. Cluster Density Visualization of documents related to topic 'Brown Adipose' queried from PubMed.

V.III. LUNG CANCER

Figure 14 shows the Co-Authorship network for PubMed documents related to the topic *Lung Cancer*. From the plot, we can observe three major clusters with prominent workers observed such as *Liu Y*, *Wang Y*, *Zhang L* primarily concentrated in the red cluster indicating rigorous work done by these authors. Clusters of other color have more uniformly spread nodes indicating steady and uniform work among the authors of those clusters.

The word network for Lung Cancer shown in Figure 15 highlights certain key terms such as cell, expression and cancer patient respectively. Each cluster provides certain insights like the blue clusters shows the regions affected due to lung cancer while the green cluster highlights the detection and treatment methods. Finally, from the item and cluster density plots shown in Figure 16 and 17, the density of terms with respect to semantically related cooccurring terms present in their neighborhood and the density of terms for the potential clusters can be observed.

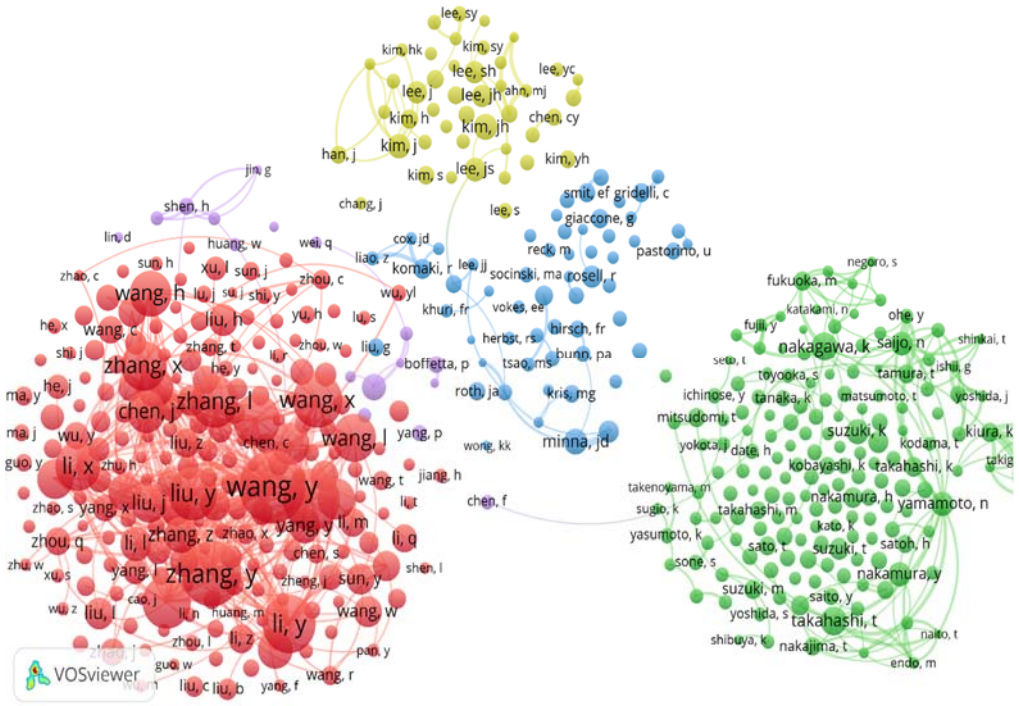


Figure 14. Co-Authorship Network Visualization of documents related to topic ‘Lung Cancer’ queried from PubMed.

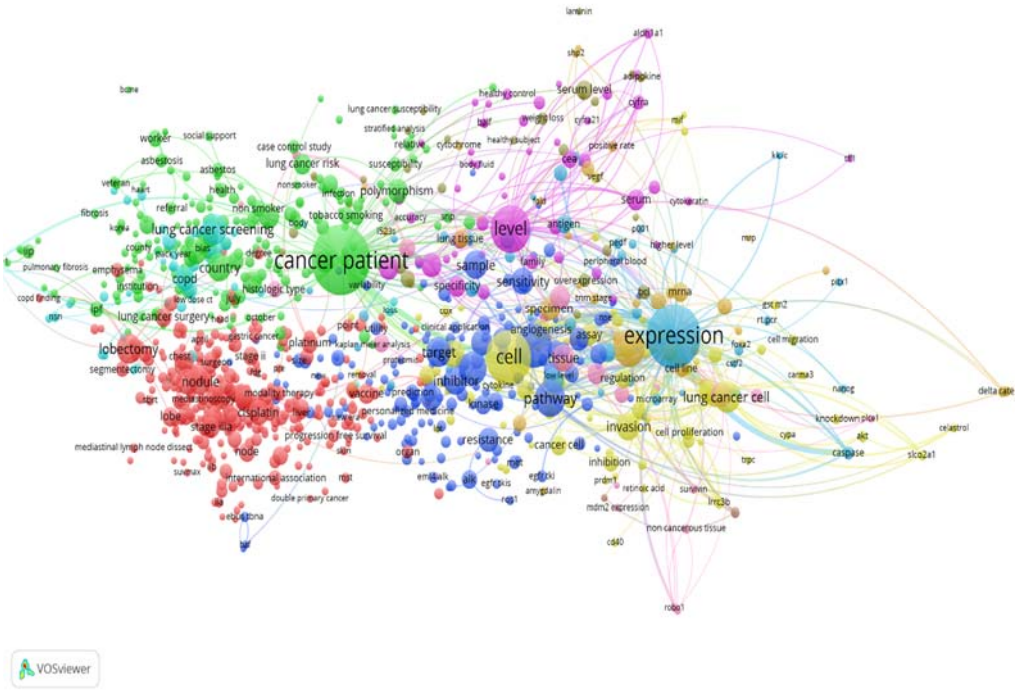


Figure 15. Co-Occurrence Word Network Visualization of documents related to topic ‘Lung Cancer’ queried from PubMed.

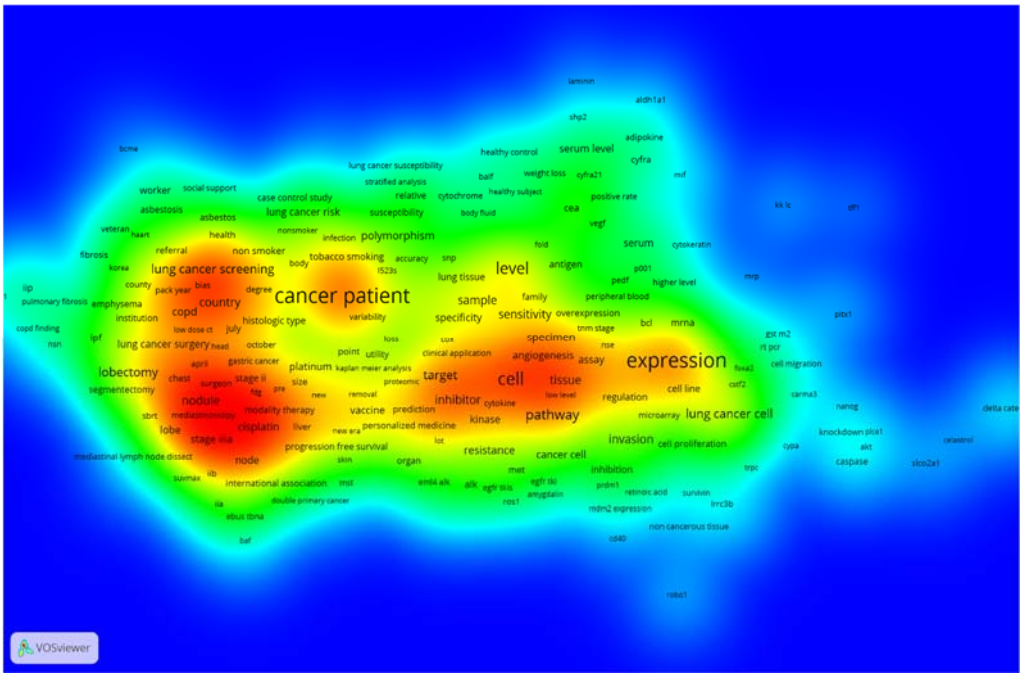


Figure 16. Item Density Visualization of documents related to topic 'Lung Cancer' queried from PubMed.

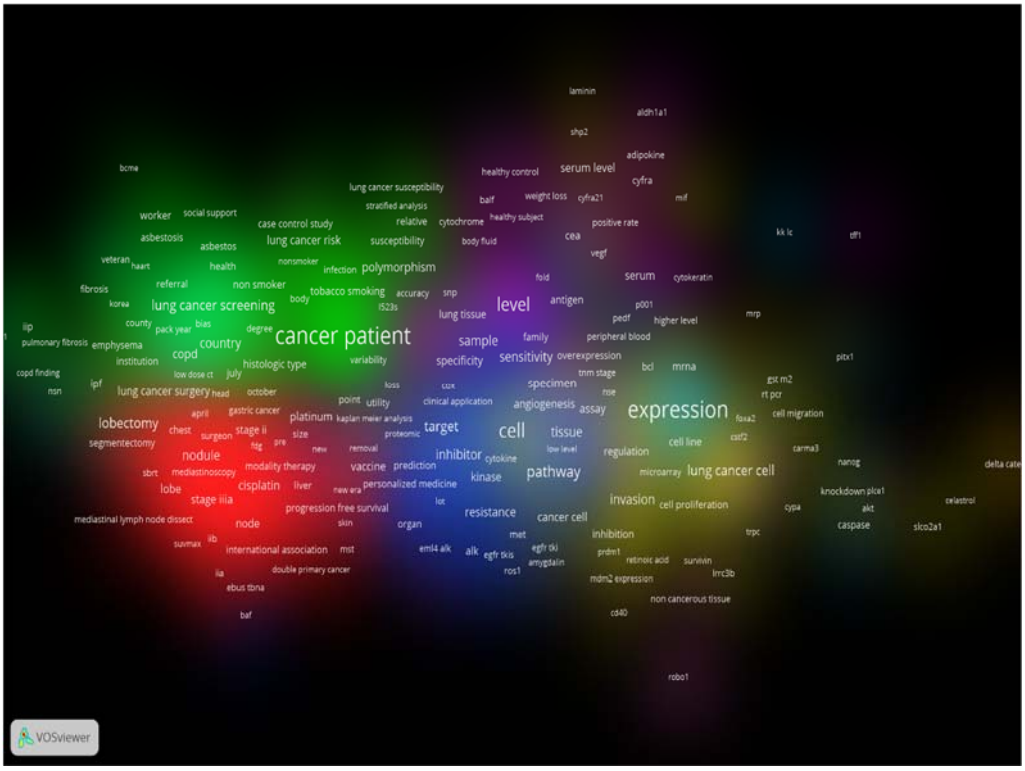


Figure 17: Cluster Density Visualization of documents related to topic 'Lung Cancer' queried from PubMed.

V.IV. TUBERCULOSIS

Figure 18 shows the Co-Authorship network for PubMed documents related to the topic *Tuberculosis*. From the figure, prominent authors can be observed based on their node sizes like, *Harries AD*, *Van Soolingen*, *Wang J*, *Narayanan PR* etc. Various clusters can also be observed shown in distinct colors with the authors in the red cluster can be seen to be widely connected with authors present in other clusters, thereby indicating prime work done by these authors related to tuberculosis.

The word co-occurrence network for Tuberculosis is shown in Figure 19. The network highlights many key terms such as certain types of tuberculosis like abdominal tuberculosis, pulmonary tuberculosis, neck tuberculosis etc. The network also lists the names of certain vaccines and resistance techniques related to tuberculosis. Finally, on deeply studying the network names of certain places like India, North Carolina, England etc can be found in association with terms like healthcare workers, survey, treatment success rate etc indicating that these places are playing a major role in spreading information and public awareness related to disease and are providing proper treatment to people affected by tuberculosis.

Finally, from the item and cluster density plots shown in Figure 20 and 21, the density of terms with respect to semantically related cooccurring terms present in their neighborhood and the density of terms for the potential clusters can be observed.

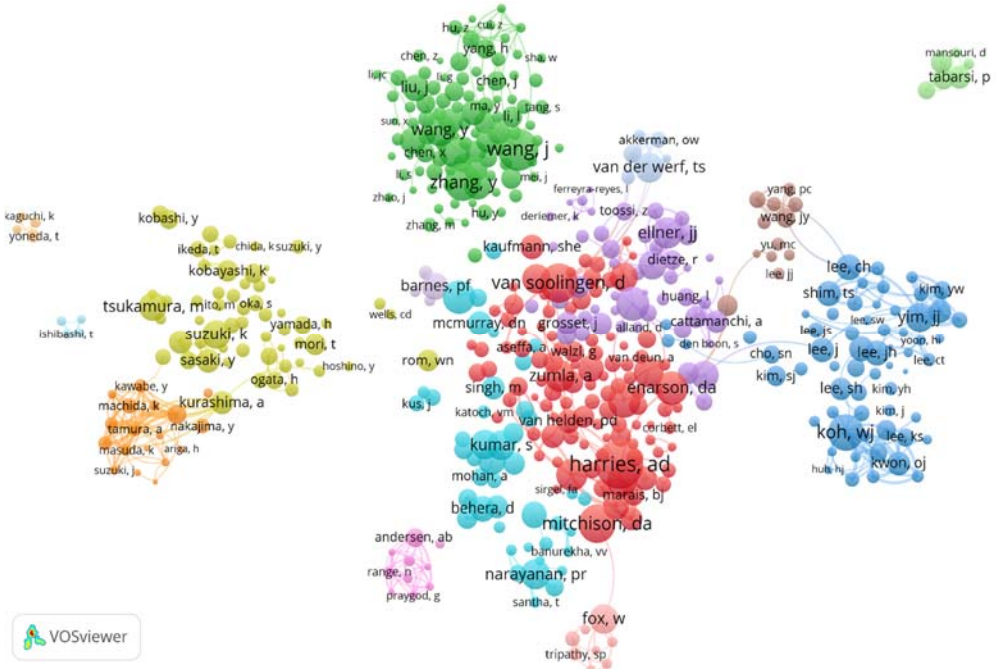


Figure 18. Co-Authorship Network Visualization of documents related to topic ‘Tuberculosis’ queried from PubMed.

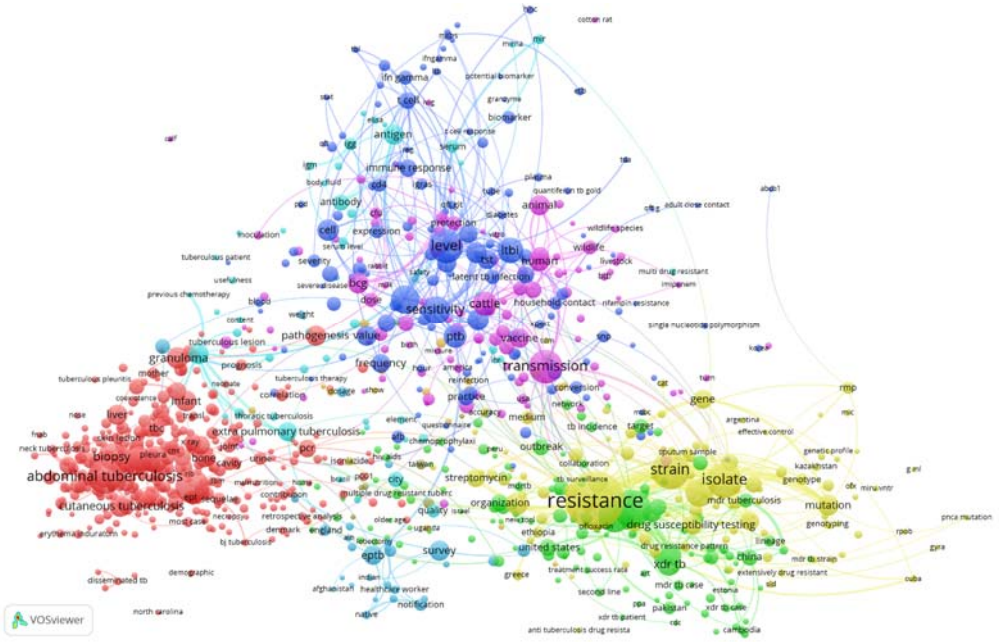


Figure 19: Co-Occurrence Word Network Visualization of documents related to topic ‘Tuberculosis’ queried from PubMed.

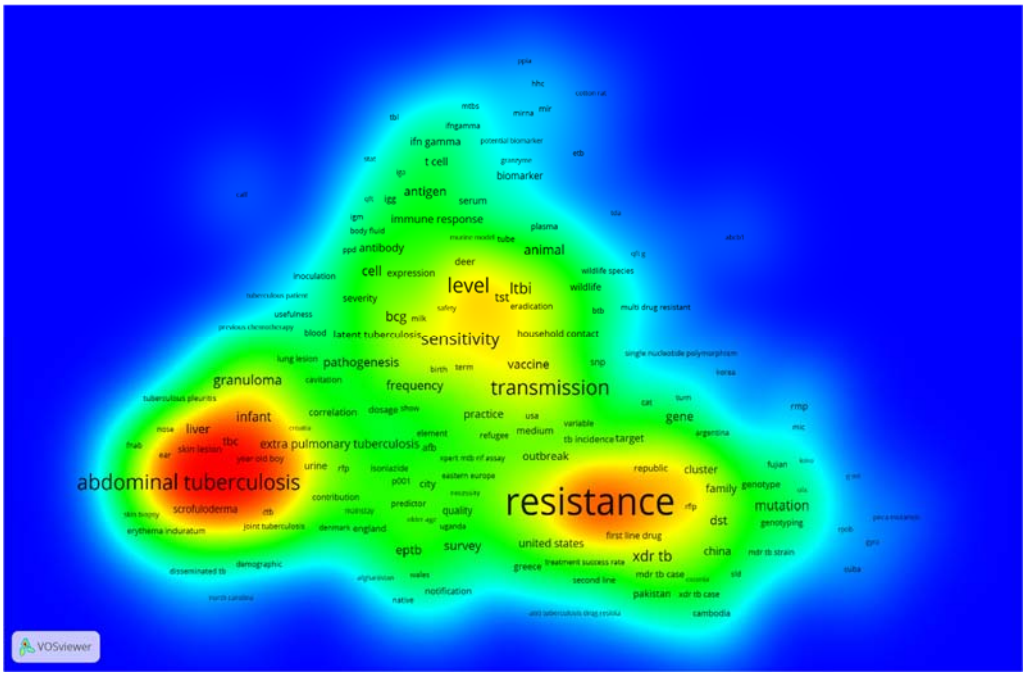


Figure 20: Item Density Visualization of documents related to topic 'Tuberculosis' queried from PubMed.

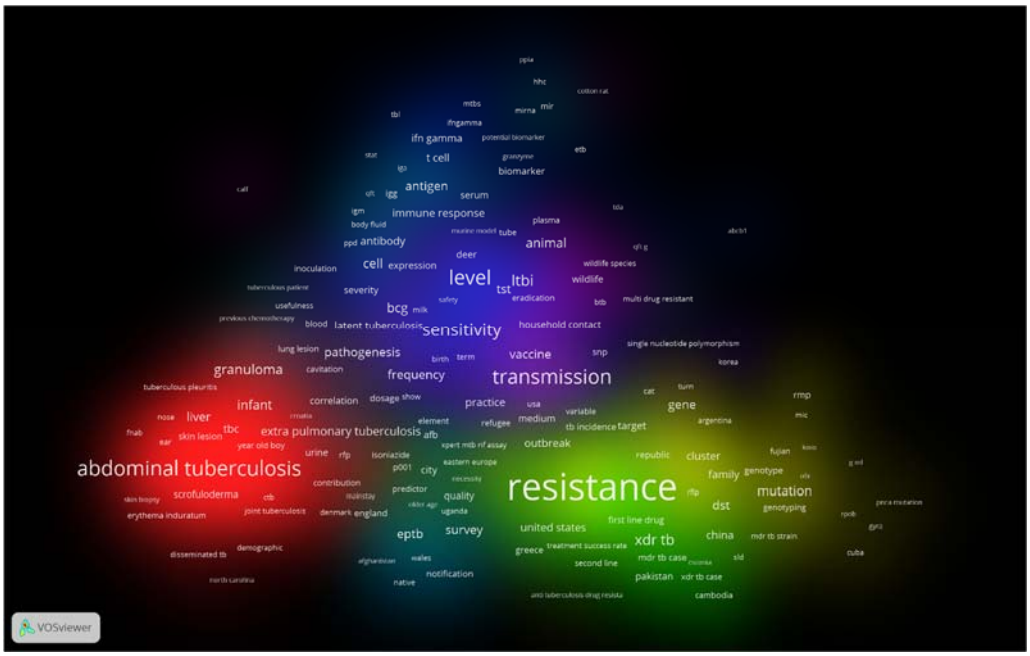


Figure 21. Cluster Density Visualization of documents related to topic 'Tuberculosis' queried from PubMed.

VI. CONCLUSION AND FURTHER RESEARCH WORK

The main motivation behind the present work is to diminish the limitation of human cognition and perception, in handling and examining enormous amounts of information. The present analytical work endeavors to exploit various data visualization tools and techniques to proffer a potential solution to overcome the problem of analyzing overwhelming amounts of information.

Although there are numerous data visualization techniques employed in thousands of domains, but the domain that majorly interested us was text. Thus, in this paper, we limited ourselves to data visualization techniques primarily concerned with lexical and text data. There are numerous techniques used for visualization of text as summarized in Figure 2. One such technique that has been used extensively in the past, especially in the analysis of research documents and case studies is Network and Citation visualization. In this paper, we studied three types of network visualizations to discover better insight into the biological sciences research papers from biomedical field. They are network visualization, the overlay visualization, and the density visualization respectively.

A survey of various NLP and Open Source based free Network and Citation Viewers has been carried out to assess their advantages for dark data analytics primarily focused around network and citation visualization. The aim was to get a fundamental understanding of the working of these tools and serve as a quick reference guide for some of them. The tools thus surveyed are Pajek, Gephi, Cytoscape, CitNet Explorer, and VOSviewer respectively. Finally, VOSviewer was used extensively for constructing and visualizing bibliometric networks.

The bibliometric networks helped in visualizing large amounts of complex bibliographic with ease. The strength of these networks is in the fact that they consume complex unstructured data and provide useful insights from it by exploiting core aspects of data and hence simplifying the further analysis. The only demerit involved with these networks is loss of information. Information is typically lost in reducing bibliographic data to its network representation. For example, when we construct word co-occurrence networks from the textual data, information on the context in which terms co-occur is lost. Similarly, for a citation network, authors citing each other can be discovered, but the reason for the citation cannot.

Loss of information also occurs due to the layout organisation of the visualization of these networks. For example, the distance between any two nodes present in the network is a representative of their relatedness, however, due to space constraints on the layout map, it is usually not possible to position the nodes such that the distance between the nodes reflects the relatedness of the nodes with perfect accuracy. Loss of information can be troublesome because it is difficult to determine the extent to which loss of information occurs and what effect does it have on the conclusions drawn from the network.

Network visualizations are most beneficial when used carefully in combination with other inferences. Finally, we discuss some ongoing and future developments in the visualization of citation networks. One important development, made possible by the availability of enormous amounts of computational resources, is the increasing attention towards visualization of large citation networks (e.g., Boyack et al., 2005; Klavans and Boyack, 2006; Skupin et al., 2013).

Another significant development is the shifting trend towards interactive visualizations. Interactive visualizations are essential when dealing with large citation networks. Static visualizations of large networks tend to be of limited use and pose many difficulties in displaying the detailed structure of a large network. Interactive visualizations, on the other hand allow large networks to be visualized and explored in full depth, from a general high-level overview to a very detailed low-level concept.

Lastly, there has been some interesting work done in dynamic visualization of citation networks, for example (Baur et al., 2002; Chen 2004, 2006). In summary, we expect a trend toward more interactive and dynamic visualizations that involve increasingly large bibliometric networks. Clearly, an exciting and highly challenging research agenda lies ahead of us.

APPENDIX [Check if required to keep or not, as it leads to plagiarism.]

In this appendix, we refer some of the "normalization, mapping, and clustering techniques" from (Van Eck and Waltman, 2017) that we utilized in our study and are used by the VOSviewer.

I. NORMALIZATION

VOSviewer uses various techniques such as association strength, fractionalization and Lin-Log modularity for normalization and analysis of citation networks. Here we discuss about one such technique called association strength normalization which we utilize in our use case. "Association strength normalization is primarily used to normalize for differences between nodes in the number of edges they have to other nodes." (Van Eck and Waltman, 2009). In mathematical terms,

Let a_{ij} denote the weight of the edge between nodes i and j, where $a_{ij} = 0$ if there is no edge between the nodes.

Then $a_{ij} = a_{ji}$, as all networks are treated as undirected networks by VOSviewer. The association strength normalization normalizes the network, so that any edge between nodes i and j has a weight given by

$$s_{ij} = \frac{2ma_{ij}}{k_i k_j}, \tag{1}$$

$$k_i = \sum_j a_{ij}, \qquad m = \frac{1}{2}\sum_i k_i, \tag{2}$$

where k_i and k_j denotes the total weights of all edges between nodes i and j and m denotes the total weight of all edges in the network.

II. MAPPING

The VOSviewer uses the mapping technique to position nodes in the network by minimizing the function

$$V(x_1, \dots, x_n) = \sum_{i<j} s_{ij} ||x_i - x_j||^2 \quad (3)$$

where s_{ij} is the similarity between nodes i and j, subject to the constraint,

$$\frac{2}{n(n-1)} \sum_{i<j} ||x_i - x_j|| = 1, \quad (4)$$

where n denotes the number of nodes in a network, x_i denotes the location of node i in a two-dimensional space, and $\vee x_i - x_j \vee$ denotes the Euclidean distances between nodes i and j.

III. CLUSTERING

Finally, the clustering technique used by VOSviewer to assign nodes to clusters is discussed here. Nodes are assigned to clusters by maximizing the following function,

$$V(c_1, \dots, c_n) = \sum_{i<j} \delta(c_i, c_j)(s_{ij} - \gamma) \quad (5)$$

where c_i denotes the cluster to which node i is assigned, $\delta(c_i, c_j)$ denotes a function that equals 1 if $c_i = c_j$ and 0 otherwise, and γ denotes a resolution parameter that determines the level of detail of the clustering. The number of clusters formed is in direct proportion to the value of γ. All the mapping and clustering techniques used by VOSviewer constitute a unified approach to mapping and clustering the nodes in a network.

REFERENCES

[1] Bastian, M., Heymann, S., and Jacomy, M., 2009, Gephi: an open source software for exploring and manipulating networks. International AAAI. https://gephi.org/

[2] Batagelj, V., and Mrvar, A., 1998, Pajek - Program for Large Network Analysis, Connections 212, 47-57, http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[3] Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I., 2010, DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks, Bioinformatics, Volume 26, Issue 22, p. 2924–2926, https://doi.org/10.1093/bioinformatics/btq538

[4] Baur, M., Benkert, M., Brandes, U., Cornelsen, S., Gaertler, M., Köpf, B., Lerner, J., and Wagner, D., 2002, Visone - Software for Visual Social Network Analysis. Proc 9th Intl Symp Graph Drawing (GD '01), LNCS. 2265, p. 463-464.

[5] Belter, C., 2012, Visualizing Networks of Scientific Research, http://www.infotoday.com/online/

[6] Boyack, K.W., Klavans, R., and Börner, K., 2005, Mapping the backbone of science. Scientometrics, 64(3), p. 351–374.

[7] Boyack, K.W., and Klavans, R., 2010, Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, 61(12), p. 2389–2404.

[8] Chen, C., 2004, Searching for intellectual turning points: Progressive knowledge domain visualization. Proceedings of the National Academy of Sciences, 101(suppl. 1), p. 5303–5310.

[9] Chen, C., 2006, CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), p. 359–377

[10] Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., and Herrera, F., 2011, Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for Information Science and Technology, 62(7), p. 1382–1402.

[11] Deep Dive, http://deepdive.stanford.edu/, Accessed on 10-14-2017.

[12] IDC, The Digital Universe, Dec 2012, https://india.emc.com/leadership/digital-universe/index.htm

[13] Jarneving, J., 2007, Bibliographic coupling and its application to research-front and other core documents. Journal of Informetrics, 1(4), p. 287–307.

[14] Ke, W., Borner, K., and Viswanath, L., 2004, Major Information Visualization Authors, Papers and Topics in the ACM Library, IEEE Symposium on Information Visualization, https://doi.org/10.1109/INFVIS.2004.45

[15] Kessler, M. M., 1963, Bibliographic coupling between scientific papers. American Documentation, 14(1), p. 10–25.

[16] Klavans, R., and Boyack, K. W., 2006, Quantitative evaluation of large maps of science. Scientometrics, 68(3), p. 475–499.

[17] Nakazawa, R., Itoh, T., and Saito, T., 2015, A Visualization of Research Papers Based on the Topics and Citation Network, 19th International Conference on Information Visualisation, Barcelona, p. 283-289. https://doi.org/10.1109/iV.2015.58

[18] NIH, U.S. National Library of Medicine, Display Formats: MEDLINE Format, https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/030_080.html (March 14, 2018)

[19] Perianes-Rodríguez, A., Olmeda-Gómez, C. & Moya-Anegón, F., 2010, Scientometrics 82: 307 p. https://doi.org/10.1007/s11192-009-0040-z

[20] PubMed Preprocessed Dataset, 2014, http://deepdive.stanford.edu/opendata/#pmc-oa-pubmed-central-open-access-subset, Accessed on 10-8-2017.

[21] PubMed, https://www.ncbi.nlm.nih.gov/pubmed/, Accessed on 10-23-2017.

[22] Roberts, R. J., 2001, PubMed Central: The GenBank of the published literature. In Proceedings of the National Academy of Sciences, p. 381–382.

[23] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T., 2003, Cytoscape: a software environment for integrated models of biomolecular interaction networks Genome Research, 13(11):2498-504.

[24] Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., and Matsushima, K., 2011, Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications, Technological Forecasting and Social Change, Volume 78, Issue 2, p. 274-282, ISSN 0040-1625, https://doi.org/10.1016/j.techfore.2010.07.006

[25] Skupin, A., Biberstine, J.R., and Börner, K., 2013, Visualizing the topical structure of the medical sciences: A self-organizing map approach, PLoS ONE, 8(3), e58779.

[26] The Apache OpenNLP, Welcome to Apache OpenNLP, Brand, http://opennlp.apache.org/

[27] Van Eck, N. J., and Waltman, L., 2009, How to normalize cooccurrence data? An analysis of some well-known similarity measures. Journal of the American Society for Information Science and Technology, 60(8), p. 1635–1651.
[28] Van Eck, N. J., and Waltman, L., 2010, Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), p. 523–538.
[29] Van Eck, N. J., and Waltman, L., 2014a, CitNetExplorer: A new software tool for analyzing and visualizing citation networks. Journal of Informetrics, 8(4), p. 802-823.
[30] Van Eck, N.J., and Waltman L., 2017, Citation-based clustering of publications using CitNetExplorer and VOSviewer, In Glaser, J., Scharnhorst, A. & Gl¨anzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of Scientometrics X(Y):XYZ, https://doi.org/10.1007/s00000-000-0000-0
[31] White, H. D., and McCain, K. W., 1998, Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. Journal of the American Society for Information Science, 49(4), p. 327–355.
[32] Wikipedia, https://en.wikipedia.org/wiki/PubMed, Accessed on 10-8-2017.