Massive Network Data Interactive Visualization

James Abello
DIMACS Center, Rutgers University
Piscataway, NJ, USA
Email: abello@dimacs.rutgers.edu

Fatima AlSaadeh
Rutgers University
Piscataway, NJ, USA
Email: fatima.alsaadeh@rutgers.edu

***Abstract*—** Network datasets are developed and collected from human interaction with the web, transportation, geographic data, web pages, and other thousands of categories. In the last two decades, these datasets became huge, containing a hundred million and billions of data where the challenge to read the vital information this data carries started to appear. As a result, new technologies, algorithms, programming languages, and techniques began to appear. In this project, we are adapting some of these techniques and algorithms to provide an interactive graph drawing system that can handle massive network datasets.

I. Project Description

This project will be using a collection of the listed authors' work to visualize various network datasets of a size up to 2^{16} edges utilizing a web application, different algorithms, and mechanisms that we will list in detail in this paper using javascript, D3, and WebGL. In addition to the two and three-dimensional graph visualizing that this project will use to represent the datasets, the goal is to make it interactive following different navigation techniques: selection, forks, zooming, and egonets. The plan to progress in this project is to start with 2^{13} edges datasets, make sure all the interactive visualizations are stable, convenient so the user will be able to extract information from it. Finally, scale up the vertices and edges size to 2^{16}.

A. Type of Users

This system is intended to be used by different types of users, starting from students, researchers, and analysts.
Example of Type of Users to Use this System :

- Social network analysis.
- Internet Security.
- Textual Analysis.
- Citation Analysis.

B. Software Tools and Libraries

The following libraries and tools were used to build this system with easy to use frontend and robust backend:

- NodeJS: an open-source, cross-platform, JavaScript run-time environment that executes JavaScript code outside of a web browser. [1]
- Three.js and WebGL: a cross-browser JavaScript library and application programming interface used to create and display animated 3D computer graphics in a web browser and it uses webGL. [2]
- D3: is a JavaScript library for producing dynamic, interactive data visualizations in web browsers.[3]
- 3D Force Graph: A web component to represent a graph data structure in a 3-dimensional space using a force-directed iterative layout and used for smaller auxiliary canvas in this system. [4]
- Bootstrap, HTML, and CSS: s a free and open-source CSS framework directed at responsive, mobile-first front-end web development. [5]

C. Project Planning and Progress

Agile software development life cycle was used for achieving the goals of this project was by dividing the work into tasks, weekly updates and feedback consideration, the Gantt Chart in Figure 1 shows the general project timeline.

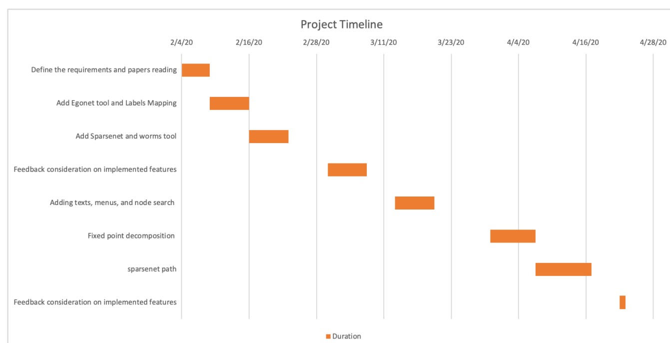

Fig. 1. Gantt Chart

II. Data, Size, Representation and Statistics

A. Data

The data this system will support is any network dataset representation, which means a collection of vertices and edges, for example, road networks, citation networks, brain networks, and social networks. Figure 2 shows the ER diagram of the system data, represented by two entities. The graph entity, which contains two main attributes, the source node id, and the target node id and the labels entity, describes the mapping label for each node id in the graph entity.

B. Data Size

The data size of the supported graphs to be visualized, and the user can navigate interactively with will be up to 2^{16} edges.

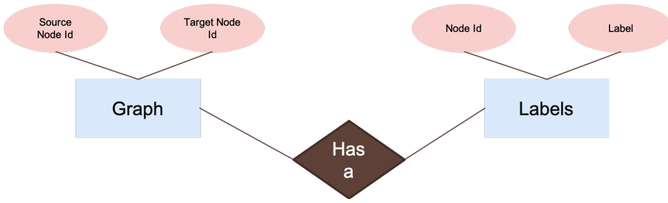

Fig. 2. ER Diagram

C. Data Representation

The data at rest will be accepted in text, CSV, and TSV files, and each record represents the source and target nodes in the graph. These files will be converted to compressed JSON files that will be stored in the computer disk. Figure 3 shows a snippet of the data for both the graph and labels entities.

node_1	node_2
9	5091
9	2951
10	1434
10	1096
10	3191

id	name
2339	ARN Colombia
4095	USAID Philippines
2850	USA.gov
5123	Spain.info
4679	HSE Ireland

Fig. 3. Graph and Labels Entities Data Snippet

D. Data Statistics

Different statistics will help to visualize such large datasets from deciding which algorithm to use for it to which is the best interactive tool.
Table I is showing an example of the statistics we are using while developing and testing this system,

statistic	example	definition
$\|V\|$	2^{13}	Number of vertices.
$\|E\|$	$2^{16} - 2^{17}$	Number of edges.
$\|cc\|$	40	Number of connected components.
$Avgdegree$	12.6	the average number of edges per node in the graph.
$Density$	0.0035	the ratio of the number of edges to the maximum possible edges

TABLE I
STATISTICS OF GOVERNMENT SOCIAL NETWORK. [6]

III. Questions To Be Answered By This System

Visualizing network data can answer many business questions and help to solve and predict real-life problems in road design, biology, internet, and network designs, and much more. In this stage of the system, it will help the user to answer the following questions in addition to insights the user can extract while interacting and visualizing the data.

- What is the best algorithm that will visualize the dataset.
- Network connected components.
- Shortest paths.
- Datasets number of vertices, edges, sparsity and density.

IV. Mode of Processing

The data will be represented using JSON files. Each file will store the connected components, vertices, edges, source and target, meta graphs, waves, waves level, layers along with summary files. The data is static and stored in the computer disk in the time of processing, for the road network example mentioned in section A, the size of total files will be around 165 MB, and this is one of the significant points to be improved which is moving the data to database storage. After the system is setup initially, if the data is changed or new data needed to be processed, the user will need to upload the latest version. Figure 4 show the general system data flow.

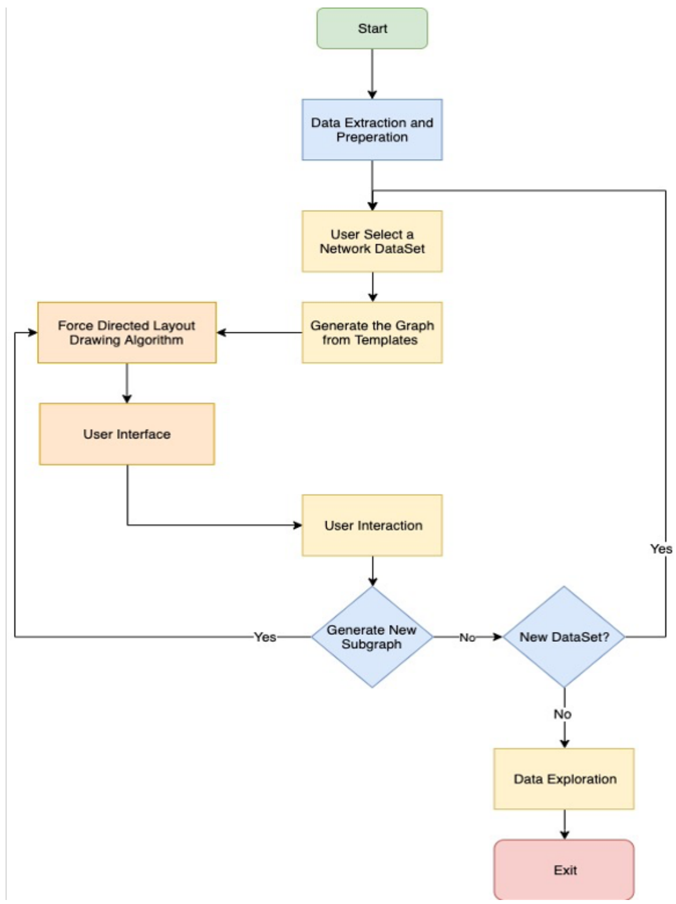

Fig. 4. System Data Flow and Processing

For the user interaction part we can break it as below:

- Key E on the keyboard will start the Egonet processing mode.
- Egonet subgraph appears on an auxiliary canvas if the size of it's edges in the range $2^10 = 2^11$.
- Key S on the keyboard will start the sparsenet navigation mode with all the related interaction mentioned in its section.
- Smooth transition between any subview and the original view, drawing a new graph, applying animation on the graph and pausing it.

V. Visual Representation and Interactivity

The main goal of this project is to visualize massive network data interactively to help extract all the information and answer many of the business questions. The data will be visualized by node-link based representation, which is an undirected graph as the main view incorporating different graph algorithms [7],[8]. Scatter plot, color, and texture will also be helping in the visualization process with other navigation tools that will responsive and take [1,3] seconds to give the results and mentioned in details below:

- Menus and Tabs that will help navigating to datasets, information about the graph, algorithms, coloring, views, and subviews. Figure 5 shows the main interface in the system.

Fig. 5. Main Interface with Menus and Tabs

- Zooming, Panning, and Selection of the nodes in the graph to make it easier for moving, browsing, and re-arranging. Figure 6

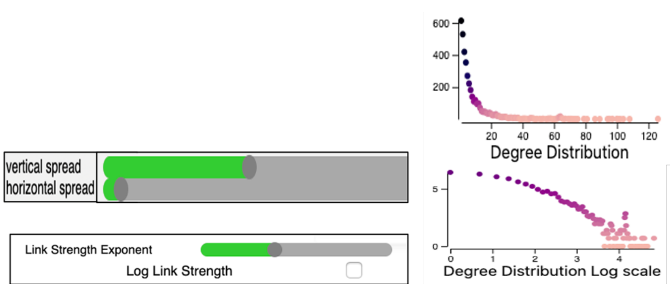

Fig. 6. h-Sliders and scatter plots

- Egonets: a subgraph induced by a vertex and its neighbor. [9] an example is shown in Figure 7. After the user press the E key on the keyboard, while brushing through the graph, the Egonet of the hovered vertex -the highlighted orange vertex in the figure - will appear in a separate view using 3D Force Graph an existing library to help drawing and to manage this view. [4]
- Sparsenet and Worms: where the sparsenet is used to show longest shortest paths in the graph, the worms are the subgraph induced by the neighbors of the vertices along these paths, Figure 8; the user can choose this tool by pressing the S key and using the h-slider to show the next or previous path, using a checkbox to show worms and the edges between these worms on demand.

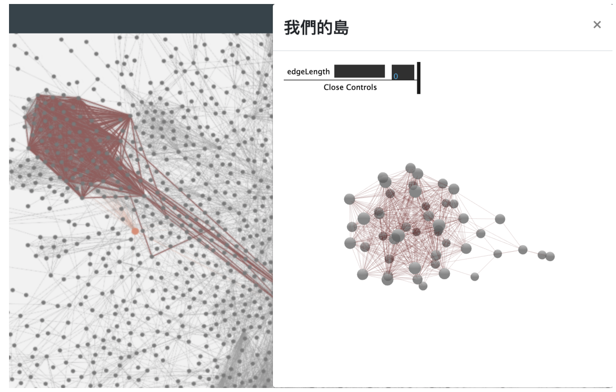

Fig. 7. Egonet

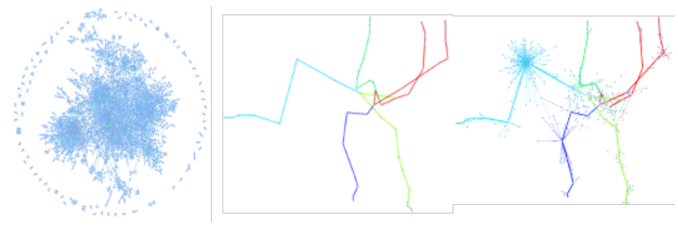

Fig. 8. The left most figure shows the original graph, the middle figure shows partial sparsenet and the right most is the partial sparsenet with the worms

- Fixed Point Decomposition: "Edge Partitioned into Maximal Edge Subgraphs of Peel value k." Where the Peel Value: "is the largest $i \in [1, deg(u)]$ such that u belongs to a subgraph of G of minimum degree i." [10] Figure 9, The user can choose this option before selecting the required dataset and can choose each layer separately using the fork described below. The 3D graph resulted in this drawing can be rotated based on the users' demand and select the speed of the rotation, increase and decrease the horizontal, vertical spread between the layers.

Fig. 9. Fixed Point Decomposition View

- Fork: to select fragments within a layer. Figure 10, after selecting a layer, the user can apply all the mentioned interactive tools on each layer chosen separately.

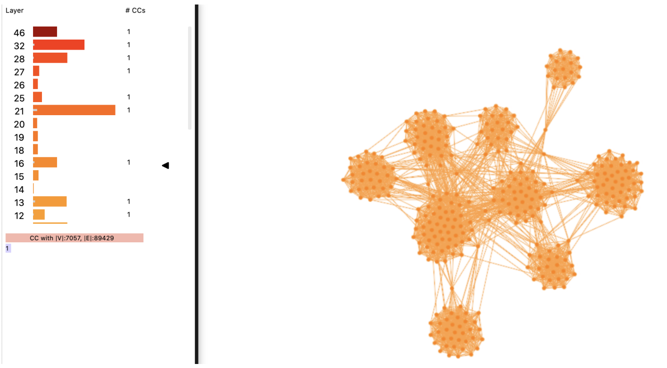

Fig. 10. Navigation to layer 16 on Figure 9 using Fork

- Sparsenet paths on Fixed Points Graph. Figure 11, The user can choose this option before selecting the required dataset and show the next paths, worms on demand.

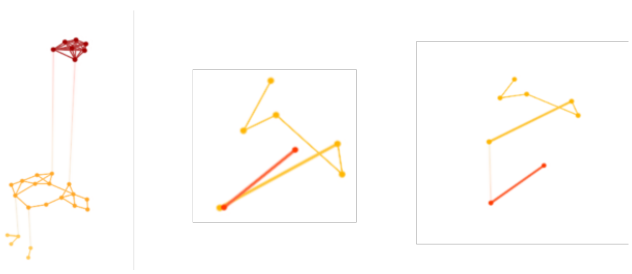

Fig. 11. Navigation to layer 16 on Figure 9 using Fork

VI. Data Insights and Interesting Findings

Different datasets were used to test and use the tools implemented in this system; In order to focus on data insights and findings, I used a TV show facebook pages where each node represents a tv show, and the edges between each pair of nodes represent mutual likes among them. [6] some interesting insights were found after visualizing the dataset and apply the interactive tools mentioned above:

- Using sparsenet paths tool the nodes along the path have a relatively similar genre, for example, tv shows of news genre were along the first path in the sparsenet
- the next path cross on the same tv show it has mutual casts among them in the following path, Constantine tv show cross with the sparsenet path start with Lost tv show that both have Jeremy Davies in the cast.
- The layers resulted from the fixed point decomposition gave the shows with the same genre and a similar rating; this applies to the clusters of the original graph drawing.

VII. Conclusion and Future Work

This system, with the combination of the algorithms, graph drawing layout, and the interactive tools, was built to be robust and straightforward for all types of users, can help to explore large graphs interactively. Extracting insights from the current version can happen manually, like finding the nodes on the same path have a relation between them. Future work of automatic semantic extraction using machine learning and data mining will be added along with improvements on the existing tools.

References

[1] Node.js. [Online]. Available: https://nodejs.org/en/
[2] Three.js. [Online]. Available: https://threejs.org/
[3] D3.js. [Online]. Available: https://d3js.org/
[4] 3d force graph. [Online]. Available: https://bl.ocks.org/vasturiano/02affe306ce445e423f992faeea13521
[5] Bootstrap. [Online]. Available: https://getbootstrap.com/
[6] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, "Gemsec: Graph embedding with self clustering," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019*. ACM, 2019, pp. 65–72.
[7] J. Abello, F. Hohman, V. Bezzam, and D. H. Chau, "Atlas: Local graph exploration in a global context."
[8] S. Dasgupta, C. H. Papadimitriou, and U. Vazirani, *Algorithms*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 2008.
[9] 3d exploration of graph layers via vertex cloning. [Online]. Available: https://www.youtube.com/watch?v=CV-0CIKg1CA
[10] J. Abello and F. Queyroi, "Network decomposition into fixed points of degree peeling."