Graph Algorithms for Visualizing High Dimensional Data

Abhinav Shankaranarayanan Venkataraman

Universitat Politecnica de Catalunya (UPC), Barcelona

27 June 2016

1. Outline
2. Project Details
3. Introduction
   - Community Structure
   - Goal of the Project
   - Planning
   - Economic Budget
   - Sustainability
4. Background
   - Graphs
5. State-of-the-art in Community Detection
6. State-of-the-art in Visualization
7. Louvain Community Detection Algorithm
   - Modularity
   - Properties of Modularity
   - Louvain Method
     - Louvain Method
     - Louvain Method
     - Louvain Method

Project Research Group

The project is done under the umbrella of LARCA(Laboratory for Relational Algorithmics, Complexity and Learning) Project Directors :

- Prof. Ricard Gavalda Mestre
- Prof. Marta Arias Vicente

What is Community?

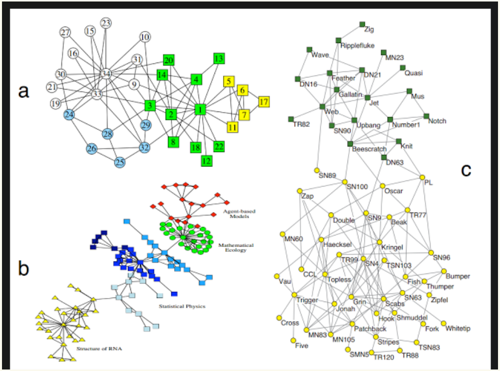

Figure : Communities: [?]

Goal of the Project

1. To survey a few algorithms that aim in community finding keeping in mind that the input is from the medical domain.
2. To choose an algorithms that benefit the purpose of organizing graphs from medical domain and for the purpose of visualization.
3. To implement the algorithms and test the efficiency of the algorithm using variety of graphs.
4. To build a Graphic User Interface (GUI) which enables visualization of the raw input on a web browser by drawing graphs.

Planning

Planning is one of the most important part of any project. In this project we divide the project into five planning phases or stages namely,

- Required knowledge acquisition
- Paper Analysis
- Design and Implementation
- Testing I
- Testing II
- Report Writing

Economic Budget

We divide the budget into 3 major categories:

- Hardware budget
- Software Budget
- Human Resource Budget

Total budget is the sum total of the three budget.
Amortized cost : Amortized cost is that accumulated portion of the recorded cost of a fixed asset that has been charged to expense through either depreciation or amortization.

Sustainability

The project is

- Economically sustainable
- Socially sustainable
- Environmentally sustainable

Graph

Theorem

A Graph G is formed by two finite sets, the set V = $\{\ v_1, v_2, \ldots, v_n\ \}$ *of vertices(also called nodes) and the set E =* $\{\ e_1, e_2, \ldots, e_n\ \}$ *of edges where each edge is a pair of vertices from V, for instance,*

$$e_i = (v_j, v_k)$$

is an edge from v_j *to* v_k *represented as G=(V,E).*

State-of-the-art in Community Detection

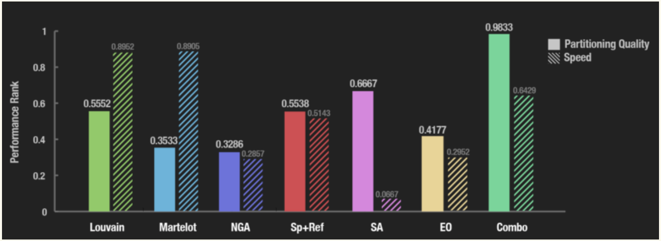

Figure : Exploring state of the art: [?]

State-of-the-art in Visualization

	Protovis.js	D3.js	Alchemy.js	Gephi
JavaScript	✓	✓	✓	
JSON Object	✓	✓	✓	
Robust		✓		✓
Less Overhead			✓	

Table : Comparing Visualization methods

Modularity

Definition

The *modularity* of a partition is a scalar value between -1 and 1 that is used to measure the density of the links inside the communities as compared to the density of the links between the communities. It is denoted by Q.
Q = (Number of Intra-Cluster Communities) - (Expected number of Edges)

Formal definition of Modularity

Definition

Formally,

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - P_{ij} \right) \delta(C_i, C_j) \tag{1}$$

$$\delta(C_i, C_j) = \begin{cases} 1, & if C_i = C_j \\ 0, & otherwise \end{cases} \tag{2}$$

The following are few properties of modularity:

- Q depends on nodes in the same clusters only.
- Larger modularity implies better Communities.
- $$Q(C_s) \leq \frac{1}{2m}\sum_{ij} A_{ij}\delta(C_i, C_j) \leq \frac{1}{2m}\sum_{ij} A_{ij} \leq 1 \tag{3}$$
- Value taken by Q can be negative
- Maximizing Modularity is considered as **NP**- Hard

Louvain Algorithm [?]

Louvain algorithms is the state of the ar community detection Algorithm. This algorithm has two phases. The diagram shows the

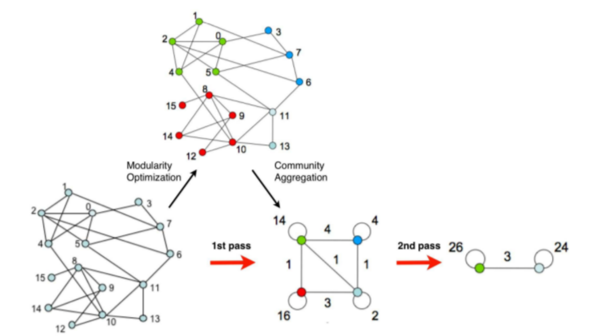

Figure : Visualization of the steps of our algorithm. This was taken from the paper ”Fast unfolding of communities in large networks” [?]

Louvain Algorithm Pseudocode

Louvain Algorithm Pseudocode:

1. Repeat until local optimum is reached:
    1. Phase1 : Split or partition the graph by optimizing modularity greedily
    2. Phase2 : Agglomerate the found clusters into new nodes

First phase in Louvain

Louvain Algorithm Pseudocodefor Phase1:

1. Assign a different community to each node.
2. For each node v_i
   - For each $v_j \in$ N(v_i),consider removing v_i from community of v_i and place it in the community of v_j
   - Choose v_i into community of neighbour that leads to highest modularity gain (Greedy Choice).
3. Repeat until no improvement can be done

Personal Learning

Since the project had more scope for exploration. My interest in Data Visualization has increased. My interest in graphs has increased. My python programming skill has also increased along with that I have also learned to code for web technologies on my own.

Software tools

1. git
2. github pages
3. Linux OS

List of References that were used

Thank you

Thank you for all those who supported me throughout the project. It was a Great time at Barcelona working with Prof.Ricard and Prof.Marta.