Universitat Politècnica de Catalunya

Bachelors in Computer Science and Engineering

Minor Heading

Graph and matrix algorithms for visualizing high dimensional data

---

Director:
Dr. Ricard Gavalda Mestre

Co-Director:
Dr. Marta Arias Vincente

Bachelors Thesis of :
Abhinav Shankaranarayanan Venkataraman

June 27,2016

To my Mother, Father, Professors and Friends. I owe a lot to My professors Ricard Gavalda and Marta Arias and to Babaji at Gurudwara

Abstract

Motivated by the problem of understanding data from the medical domain, we consider algorithms for visually representing highly dimensional data so that "similar" entities appear close together. We will study, implement and compare several algorithms based on graph and on matrix representation of the data. The first kind are known as "community detection" algorithms, the second kind as "clustering" algorithms. The implementations should be robust, scalable, and provide a visually appealing representation of the main structures in the data.

Acknowledgement

I would like to Acknowledge the support provided by my faculty and admins at my home university – SASTRA University, Thanjavur and UPC Barcelona for supporting me throughout the project.

Contents

1 Introduction

In this section we provide an overview of the entire work. We mention the context of the project we have studied, approaches that we have used, goal of the project. We also provide the intended planning, economic estimate and sustainability of the work that has been done.

1.1 Context Of the Project

In the present day scenario, the modern science of algorithms and graph theory has brought significant advances to our understanding of complex data. Many complex systems are representable in the form of graphs. Graphs have time and again been used to represent real world networks. One of the most pertinent feature of graphs representing real system is community structures or otherwise known as clusters. Community can be defined as the organization of vertices in groups or clusters, with many edges joining the vertices of the same cluster and comparatively fewer vertices joining the vertices in another neighbouring cluster. Such communities form an independent compartment of a graph exhibiting similar role. Thus, Community detection is the key for understanding the structure of complex graphs, and ultimately educe information from them.

1.2 Approaches

In this section we discuss the various approches that are involved in dealing with the input to the project for community identification, for clustering and for visualization purposes.

1.2.1 For Community Identification

Virtually in every scientific field dealing with empirical data, primary approach to get a first impression on the data is by trying to identify groups having "similar" behaviour in data. There are numerous methods to achieve this objective of which

- Community Detection
- Clustering

1.2.1.1 Community Detection

1.2.1.2 Definition of a Community Communities are a part of the graph that has fewer ties with the rest of the system. Community detection traditionally focuses on the graph structures while clustering algorithms focuses on node attributes.

1.2.2 Clustering

Traditional Clustering Methods are as follows:

- Graph Partitioning
- Hierarchical Clustering
- Partitional Clustering
- Spectral Clustering

1.2.2.1 Graph Partitioning The problem of graph partitioning is a hard problem. The graph partitioning consists of dividing the vertices in g groups of predefined sizes. There are two broad categories of methods : Local and Global.

1.2.2.2 Hierarchical Clustering Hierarchical clustering aims to identify groups of vertices with high similarities. It can be calssifies into two categories:

1. *Aggolomerative algorithm* : in one in which Agglomerative algorithms, in which clusters are iteratively merged if their similarity is sufficiently high
2. *Divisive algorithms*, in which clusters are iteratively split by removing edges connecting vertices with low similarity.

1.2.2.3 Partitional Clustering

1.2.2.4 Spectral Clustering

1.2.3 For Visualization

Graph visualization is a important task in various scientific application. Visualization of data as graphs Visualizing these data as graphs provides the non-experts with an intuitive means to explore the content of the data, identify interesting patterns, etc. Such operations require interactive visualizations (as opposed to a static image) in which graph elements are rendered as distinct visual objects; e.g., DOM objects in a web browser. This way, the user can manipulate the graph directly from the UI, e.g., click on a node or an edge to get additional information (metadata), highlight parts of the graph, etc. Given that graphs in many real-world scenarios are huge, the aforementioned visualizations pose significant technical challenges from a data management perspective

1.2.3.1 Web UI and Time per Query

1. *Program execution and Computation* :
2. *Building the JSON Object*
3. *Communication Time*
4. *Rendering*

The total time is the sum of all the above times.

1.2.4 Computational Complexity

The estimate of the amount of resources required for by the algorithm to perform a task is defined as computational complexity. The humongous amount of data on the real graphs or real networks that are available in the current scenario causes the efficiency of the clustering algorithm to be crucial.

In a brief, Algorithms that have polynomial complexity describe the Class **P**. Problems whose solutions can be verified in a polynomial time span the class **NP** of *non–deterministic polynomial time* problems, which includes **P**. problem is **NP**-hard if a solution for it can be translated into a solution for any **NP**-problem. However, a **NP**-hard problem needs not be in the class **NP**. If it does belong to **NP** it is called **NP**-complete. The class of **NP**-complete problems has drawn a special attention in computer science, as it includes many famous problems like the Travelling Salesman, Boolean

Satisfiability (**SAT**), Linear Programming, etc. The fact that **NP** problems have a solution which is verifiable in polynomial time does not mean that NP problems have polynomial complexity, i. e., that they are in **P**. In fact, the question of whether **NP**=**P** is the most important open problem in theoretical computer science. **NP**-hard problems need not be in **NP** (in which case they would be **NP**-complete), but they are at least as hard as **NP**-complete problems, so they are unlikely to have polynomial complexity, although a proof of that is still missing.

Many clustering Algorithms or problems related to clustering are **NP**-hard. This makes it irrelavant to use the exact algorithm, in which case we use an approximation algorithm. Approximation algorithm are methods that do not deliver the exact solution but an approximate solution but with an advantage of lower complexity. [Community detection paper]

1.3 Goal of the Project

1.4 Planning

1.4.1 Task Description

The tasks for the project have been subdivided into various task phases which are enumerated below :

- **Required knowledge acquisition**
  Before any immersion into the real topic, it was necessary to acquire the knowledge necessary to understand the problem. In this phase we familiarize with the term modularity, Louvain algorithm for community detection and various other algorithms used for community detection. Acquisition of knowledge about visualization tools to be used and make conversant with python is also required.

- **Paper Analysis**
  In this phase we analyze and compare several works about community detection and clustering algorithm over high dimensional graph-like data. Doing this we became conscious of functionalities that our proposal should have and we are thus able to guide all the subsequent phases.

- **Design and Implementation**

In this phase the project is designed and coded implementing all the functionalities of the solution.

- **Testing I**
  In this phase we test the program in order to identify errors in the implementation. It includes the successive recoding.

- **Testing II**
  In this phase we perform tests over synthetic and real data streams. We evaluate the performance of the program and we study the effects of concept drift.

- **Report Writing**
  In this phase the report of the project is written.

1.5 Economic Budget

1.5.1 An Introduction to Economic Budget

Economic management is primarily based on an estimate of income and expenditure called as budget. Development of a sustainable budget leads to proper economic management of the project. Budget and sustainability is one of the most important phase of the project management. In this phase we analyze the budget for the project. We also aim at providing an estimate of the project budget and optimize the same. We look at the expenditure from various aspects such as software costs, hardware costs, license costs and human resource costs. Additionally we also account the software for its sustainability. One important factor to note is that the budget that we describe in this section is subject to change and it may increase depending on the unexpected obstacles that we may face. For an instance when we don't get the expected results with a particular software we may have to go in for another software that may incur extra installation and operational charges.

1.5.2 Estimation of Economic Budget

We divide the overall expenditure into three categories namely hardware, software and human resources. One very important factor that we need to consider is that we only get an estimate of the total cost. This may vary depending on the systems in use. To calculate the amortization we consider to factors namely, first the overall life of the hardware or software in use. Second that the project is completed in 5 months. Hence the amortization cost comes one eighth of the actual life of the component.

1.5.2.1 Hardware Budget Hardware budget accounts for the actual and the amortized costs of the hardware elements used by the project. The cost is fictitious as it has not been developed commercially. Table 1 intents to estimate the economic cost of each of the hardware component of the project.

Table 1 - Hardware Budget

Sno:	Hardware Component	Useful Life	Total Cost(in €)	Amortized Cost(in €)
1	PC System	4	1000€	125 €
	Total		**1000€**	**125 €**

1.5.2.2 Software Budget The software budget shows an estimate for the various software used in the project along with the estimate of the software costs. It is a myth that the software doesn't get old with time just as a software gets but it wears out with time. Thus for every software there is a fixed time during which it gives maximum performance. In addition freeware software and open source software incur no cost. The cost is fictitious as it has not been developed commercially. Table 2 intents to estimate the economic cost of each of the software component of the project.

Table 2 - Software Budget

Sno:	Software Component	Useful Life	Total Cost(in €)	Amortized Cost(in €)
1	Linux OS	5	0€	0 €
2	JavaScript Engine	1	0€	0 €
3	Python Components	1	0€	0 €
4	Web.py	1	0€	0 €
5	TexMaker	1	0€	0 €
	Total		**0€**	**0 €**

1.5.2.3 Human Resource Budget The human resource budget deals with the overall expenditure spent on human resources. Every phase of the project has a cost associated with it in per hour calculation. The cost is fictitious as it has not been developed commercially. Table 3 intents to estimate the economic cost of each of the phases of the project. The cost per hour is intended as an approximation of the current cost per work hour of young analysts and developers in our environment.

Table 3 - Human Resource Budget					
Sno:	Phase	Deadline	Hours	Cost(per hour in €)	Total(in €)
1	Required Knowledge Acqusition	1 Mar 2016	70	15€/h	1050 €
2	Paper Analysis	1 Apr 2016	150	15€/h	2250 €
3	Design and Implementation	30 Apr 2016	230	20€/h	4600 €
4	Testing I	15 May 2016	75	15€/h	0 1125€
5	Testing II	31 May 2016	75	15€/h	0 1125€
6	Report Writing	15 Jun 2016	100	15€/h	1500€
	Total		**600**		**10525 €**

1.5.2.4 Total Budget The following table, Table 4, summarizes the total budget for the project. This encompasses the hardware, software and human resources budget.

Table 4 - Total Budget		
Sno:	Resource	Total Cost(in €)
1	Hardware Budget	1000 €
2	Software Budget	0 €
3	Software Budget	10525 €
	Total	**11525 €**

1.6 Sustainability

Sustainability is a key factor in any project design. We evaluate the project based on three factors of sustainability namely economic sustainability, social sustainability and environmental sustainability.

1.6.1 Economic Sustainability

In this document we specify the budget estimation of the project. From our estimation it can be said that this will be the maximum bound on the budget for the project. This takes into account all the factors namely the hardware costs, software costs and human resource costs. The cost estimated in the project is the least possible cost and hence is a nonpareil project estimate for any indistinguishable project. The budget may exceed our calculations only during unexpected times. When the proposed plan is precisely followed the estimated lower costs gets achieved. Also the product that we aim at developing here is tested with all kinds of data and we aim at building a very high quality software which in turn provides a durable software that will not wear out easily. Most of the software used in the project is open source which has zero product cost. The The hardware required is nothing but computers that becomes a mandatory part of any project in the present days.

1.6.2 Social Sustainability

The project aims at developing web based platform to perform learning cum visualization analytics. This is indirectly going to analyze the learning characteristics of the patients and provide a feedback both to the medical analyzer and health planner. This is going to improve the quality of health analysis in the state. All this requires is a simple computer connected to the internet. This has very keen social motive and this project when completed is going to improve the standard of learning in schools. Thus this has a great social responsibility. This is in turn justifies why this project has a great social sustainability.

1.6.3 Environmental Sustainability

From the sections of temporal planning and the budget planning we understand that we have a computer running throughout the project. If we make an assumption that the amount of energy used by a single computer comes to around 250 watts. And given that we spend 500 hours on the project then the energy expended is 125KW. This amounts to 48.125 kg of CO_2 . This is indeed a high amount but well within the permissible limits. This can be reduced by reducing the code size which is possible by reusing the already existing code. But the project is actually environmentally sustainable.

2 Background Knowledge

In this section we present the background knowledge required to understand and solve the problem

2.1 Graph Notation

Graph ,G , is construct consisting of two finite sets, the set $V = \{ v_1, v_2, \ldots, v_n \}$ of vertices 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). In other words E $\subset V^2$, which is the set of all unordered.

2.2 Graph Matrix Notation

Graphs can be appropritely represented in the form of matrices for instance, adjacency matrix, admittance matrix etc.,
Let G=(V,E) be a simple graph with vertex set **V** and edge set **E**, then the adjacency matrix is square $|V|^2$ matrix **M** such that its element $M_{i,j}$ is *1* when there is an edge from v_i to v_j,where $v_i \in$ **V** , $v_j \in$ **V** and *0* when there is no edge. The adjacency matrix of a graph of order n entitles the entire the topology of a graph. The diagonal elements of the adjacency matrix are all *0* for undirected graphs **M**.

The sum of the elements of i-th row or column yields the degree of node i. If the edges are weighted, one defines the weight matrix **W**, whose element W_{ij} expresses the weight of the edges between vertices i and j.

The *spectrum* of a graph **G** is the set of eigenvalues of it's adjacency matrix **M**. If **D** is the diagonal matrix whose element $D_{\,i,i}$ equals the degree of vertex i ($v_i \in V$).

2.3 State-of-the-art in Community finding

2.4 State-of-the-art in Clustering

2.5 State-of-the-art in Graph Visualization

explain technical concepts in more detail. for exampleequivalence of graph and matrix representations. state-of-the art in community finding, and clustering state-of-the art in graph visualization

3 Community Finding Algorithm

3.1 Louvain Algorithm

3.1.1 Introduction

3.1.2 Reasoning

3.1.3 Description

3.1.4 Implementation

3.1.5 Experiments

3.1.6 Result

4 Matrix Based Algorithm

4.1 Matrix Algorithm

4.1.1 Introduction

4.1.2 Reasoning

4.1.3 Description

4.1.4 Implementation

4.1.5 Experiments

4.1.6 Result

5 Visualization

5.1 Alchemy.js

5.1.1 Introduction

5.1.2 Reasoning

5.1.3 Description

5.1.4 Methods and Library

5.1.5 Result

6 Overall System Description

6.1 Alchemy.js

6.1.1 Introduction

6.1.2 Implementation Benefits

sdgbf akjsfkldsnfksndafnadfa fasfadf adfasfasfadf asfasfadf afadfadf

6.1.3 Description

6.1.4 Result

6.2 Benefits to the community

This can be used in places where there is difficulty in visualization of a very complex landscape of data such as medical domain. In Medical domain a patient can be a vector of diseases and visualization of such patients (patients graph–which shows relations of how two patients are similar, a graph in which patient-patient edge weight is the similarity value) would be useful for analyzing and predicting the disease landscape of a region and in turn multiple regions.

7 Conclusion

7.1 Goals Achieved

7.2 Revision of Planning and Budget

7.3 Future Works

7.4 Personal Conclusion

This is one of the greatest project experience.