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


UPC

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

1.3 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

1.3.1 Community Detection

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

1.3.2 Clustering

Traditional Clustering Methods are as follows:

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

1.3.3 For Visualization

1.3.4 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.

community finding and clustering. separately, visualization tools

1.4 Goal of the Project

1.5 Planning

1.5.1 Task Description

- **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 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.

1.6 Budget

1.7 Sustainability

2 Background

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).

2.2 Matrix Notation

2.3 Equivalence between Graph and Matrix Represenation

2.4 State-of-the-art in Community finding

2.5 State-of-the-art in Clustering

2.6 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