Graph based clustering

Author Disambiguation
Data Preprocessing
DMKM

Carlos López Roa
me@mr3m.me

April 15, 2016

Abstract

As part of the author disambiguation a feature construction based on graph theory is desired to exploit the inner structure of the data as a social network.

Introduction

Collaboration Networks first described by Paul Erdös consists of a graph of individuals represented as dots and they relations represented as edges connecting the dots.

Collaboration Networks can be fully defined with two sets, a set of vertices and a set of edges. In this case we can define the set of vertices as the set of signatures composed of $(id, article - id, author - id)$ where id is a consecutive unique number assigned to each signature, $article - id$ is the unique number assigned to each article and $author - id$ is the unique number assigned to each different author string. And the set of edges as a binary relation (undirected) between two signatures if they have the same $article-id$ and different $author - id$, the last to avoid *self relations*.

The described graph can be proved to be *simple*[1] and *small world*[2], the later characterises the *distance* between any two points, defined as

$$d(v_i, v_j) = \arg\min_k P(v_i, v_j), \tag{1}$$

that is the minimum length path between to vertices[3], to be

$$\begin{aligned} E[d(v_i, v_j)] &= K \\ K &\propto \log N, \end{aligned} \tag{2}$$

that is the average distance between any two vertices grows as the logarithm of the number of vertices.

In some cases a collaboration can be characterised as *scale free*[4] if the presence of hubs the average distance grows as

$$K \propto \log\log N. \tag{3}$$

Small world graphs can be modelled as *Watss-Strogatz random graphs* with just two parameters, namely the clustering coefficient and the average node-to-node distance as shown in 1

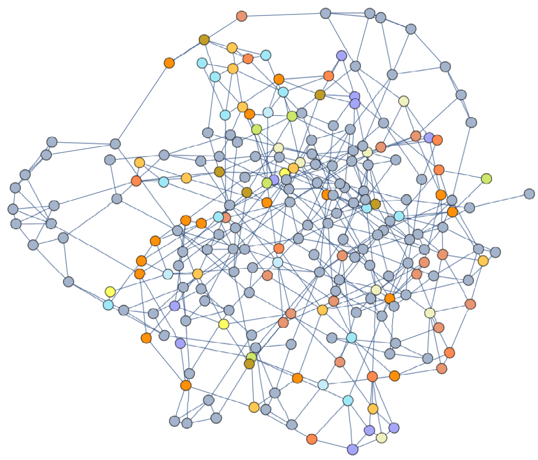

Figure 1: Watss-Strogatz random graph with 15^2 nodes.

Methods

The objective is to construct a feature based on the distance between two signatures as part of the analysis and with this help to decide whether two signatures correspond to a particular community or cluster.

Given the data set $\{(s_i, s_j) \in S \times S, i \neq j\}$ where S is the set of signatures. We can compute the distance $d(s_i, s_j)$ for each pair by first constructing the graph at depth λ between s_i, s_j, that is to take the neighbours of s_i and the neighbours of the neighbours of s_i and so on λ times, and the same for s_j. With this

[1] With no self relations, all edges are undirected thus making the edges a set and not a multiset.

[2] In which most nodes are not neighbors of one another, but most nodes can be reached from every other node by a small number of hops or steps

[3] If the graph is not *connected* and there exists no path between to vertices, their distance is said to be infinite

[4] or ultra-small world

graph we need to find the minimum length path. If there exists no path, then we can assume that the distance $d(s_i, s_j)$ is at least bigger than 2λ and include this distance as feature in the dataset.

Results

A graph with $5,000$ edges and nodes was constructed as described, and plotted as show in figure 2

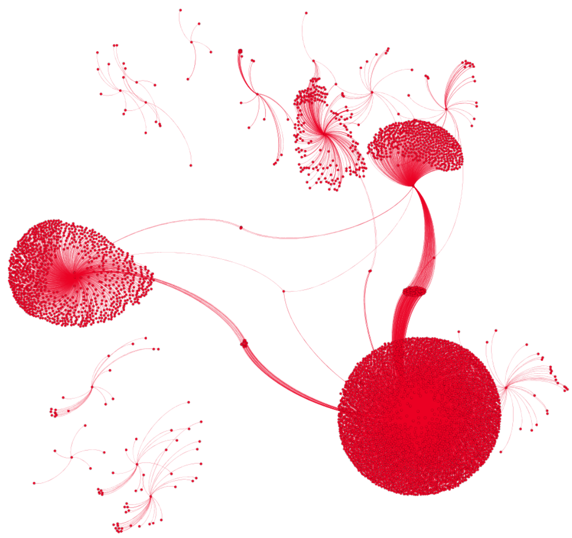

Figure 2: Graphical result of the incomplete graph of around 5,000 nodes and edges.

This result was produced using a brute-force approach in SQL. A more refined function is yet to be implemented as a recursive call which will reduce the complexity. The complete graph can be of size 10^12 which is unfeasible.