|  |
| --- |
|  |
| Network Analysis of UFO Sightings |
| Programming Assignment (Social Networking Analysis) |
|  |
| **Chetan Konaje** |
| **4/24/2013** |

|  |
| --- |
| This is attempt to apply network analysis principles to data captured by UFO sightings and analyze the characteristics of the data and whether it points to certain facts. |

Network Analysis of UFO Sightings

UFO sightings are one of interesting phenomena’s that no one on this planet can explain with certainty. There are various organizations like SETI which are on the hunt of extra terrestrial life but public knowledge of this has been limited. Various UFO aficionados have put together information on internet to capture these evidences and provide commentary on this subject. There are theories of UFO sightings linked to disasters which could be preposterous but worth understanding. If there is belief in their existence, then believers have thought these civilizations may be far advanced then ours. Therefore we can dwell on various observations as see what immediate affect has been seen to analyse this data.

Therefore it could be worthwhile to see how the sightings and other data like disasters can be correlated using network analysis to provide thought provoking perspectives.

It should be seen whether there is correlation of sightings on where the UFO was seen? Or How many times UFO’s were seen at location? Or On particular date why there were many sightings? Any reasons why they revisit some popular locations like Phoenix, AZ.

Can we tie these sightings with other relevant data to establish any relationships? Interesting???

The approach here is to create network of sighting and location where the UFO has been seen. This may provide some clues as to which are the locations where large number of sightings have been seen.

Another approach is to club the sightings and network them to their locations. This could provide us data of virtual sighting communities that are existing and how are they spread. Are they in same location or different? Unfortunately the data does not consist of more attributes like time of sighting so that we can club these together to see more relations.

# Data Preparation

The data for these UFO sightings are available here <http://www.infochimps.com/datasets/60000-documented-ufo-sightings-with-text-descriptions-and-metada> . I later saw this website <http://www.nuforc.org/webreports.html> which was more useful but for want of time did work with previous one. I have written two python programs which were used to create ***graphml*** file which contains the necessary nodes and edges. In the first program crtdata.py have created edges between each sighting and node, while in the second program crtdata1.py clubbed the sightings as per date and created the necessary edges which reduced the number of nodes. To reduce the data scope, have used data that was more than Jan 2005.

Steps to run the python program

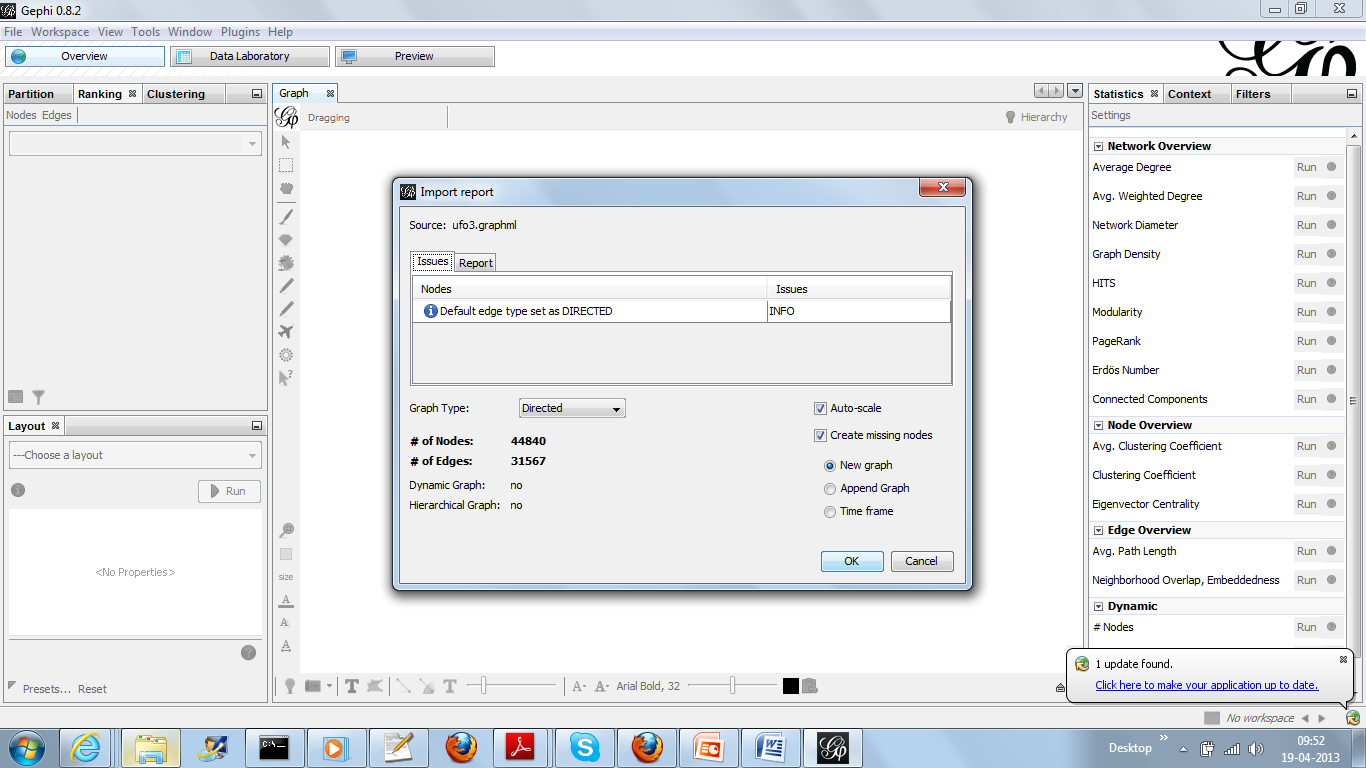
1. Install **python**
2. From command window (if on windows)

>**python crtdata.py**

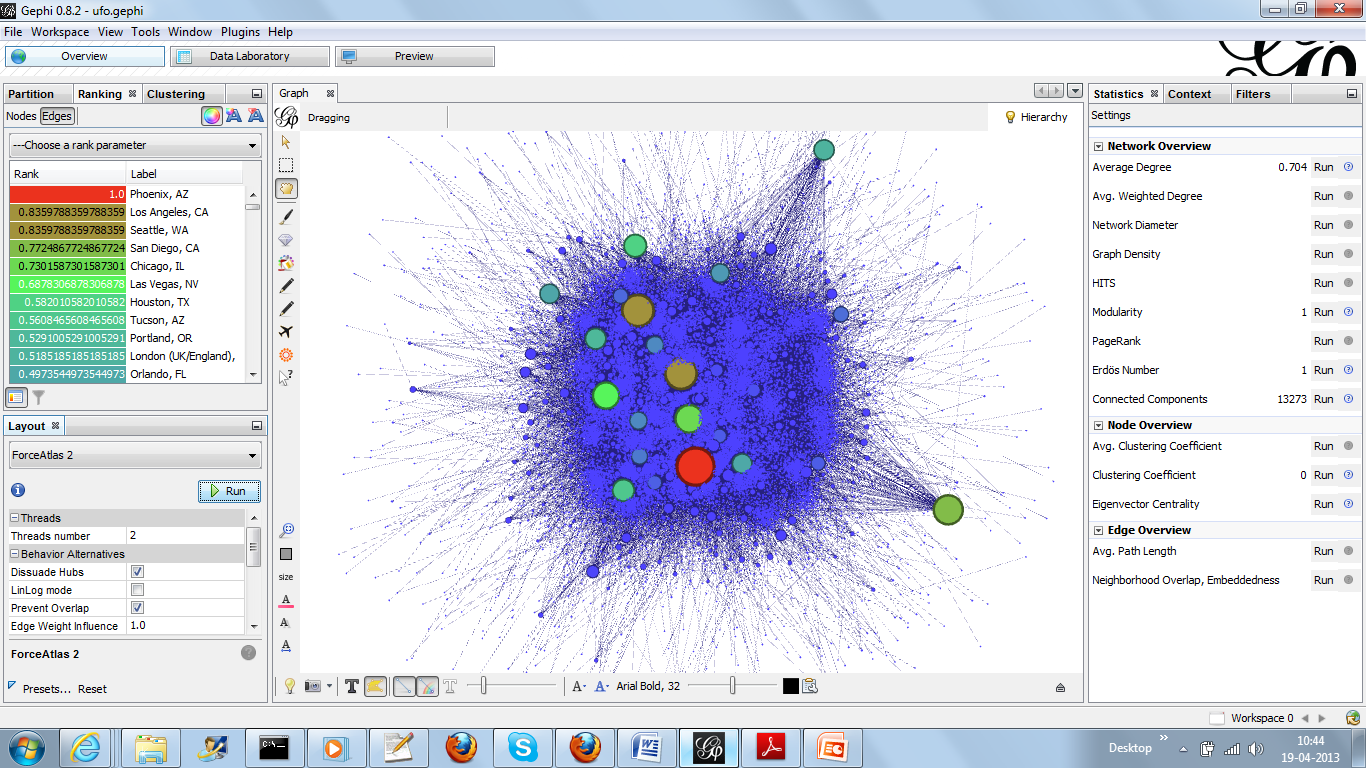
3) The script reads the **ufo.csv** file and creates **ufo.graphml** file which can be opened using **gephi** tool.

# Approach 1

In this approach, we have one edge created per sighting date to the location.



Degree Centrality indicates Phoenix node has the largest number of degrees and maximum of sightings have been at Phoenix. Most of the centrality provides not much information as the network has not been properly formed. Eigenvector Centrality does provide highest ranking for Phoenix.

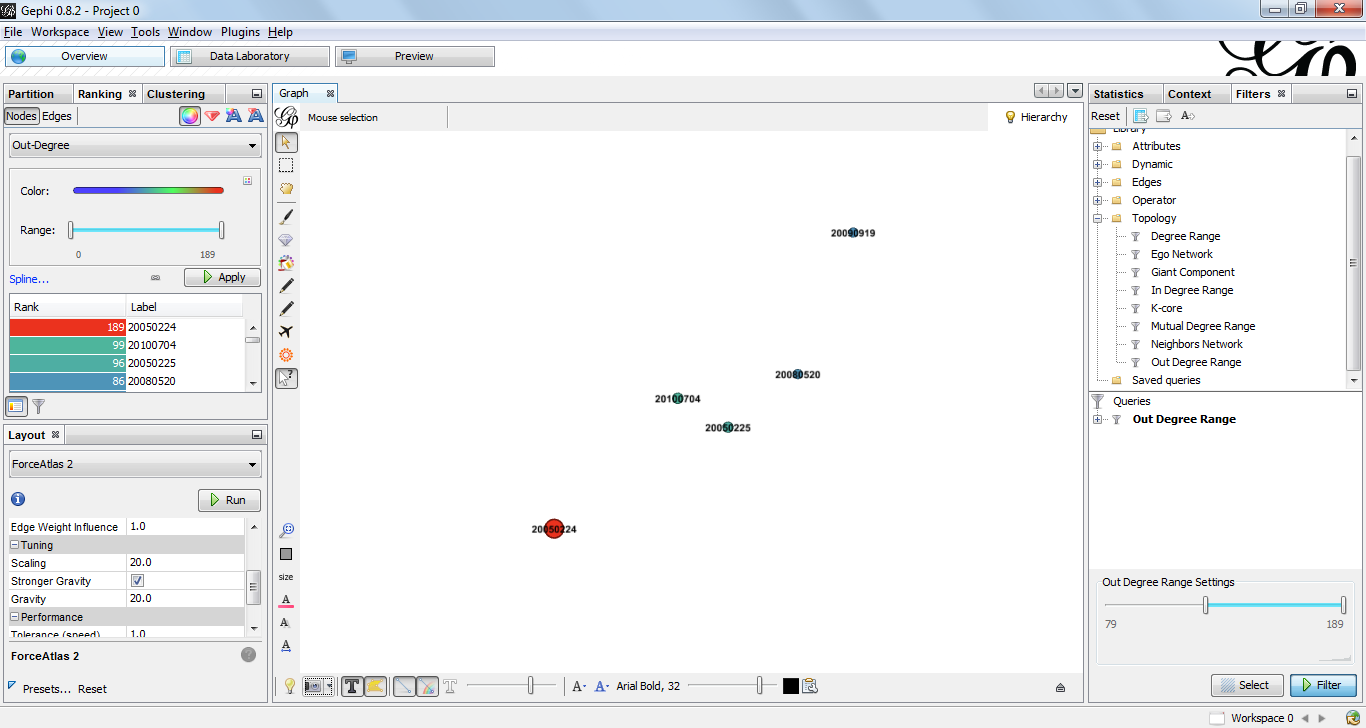


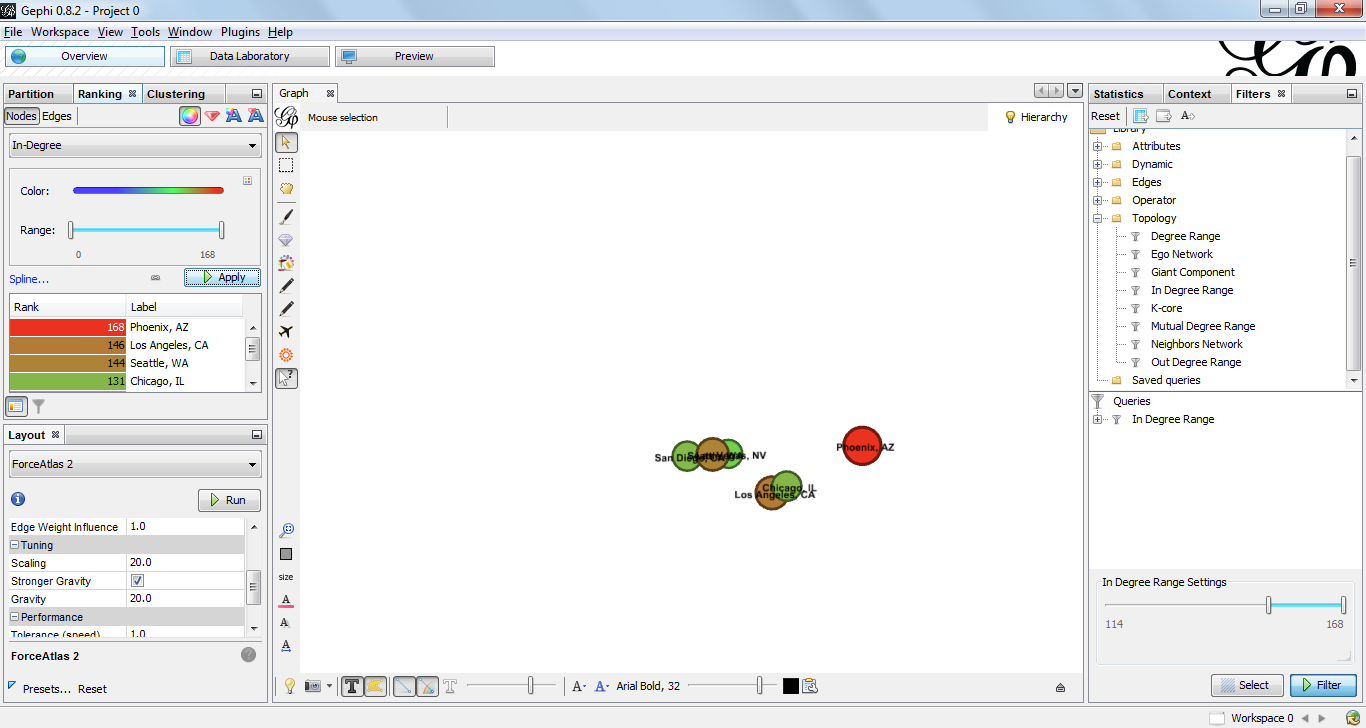
Highest out degree indicates highest sightings on particular date. From the data, it’s clear that it is 24th February 2005 which has 824 sightings! As most of the network has been sparse and to bring an element of network, approach 2 is adopted which clubs the sightings with common date. This ensures that the network has balanced number of in-degree and out-degree.

# Approach 2

Now construct the graph in such way that we club the sightings for particular date so now the out-degree of sighting nodes will be greater than 1 in most of the cases.

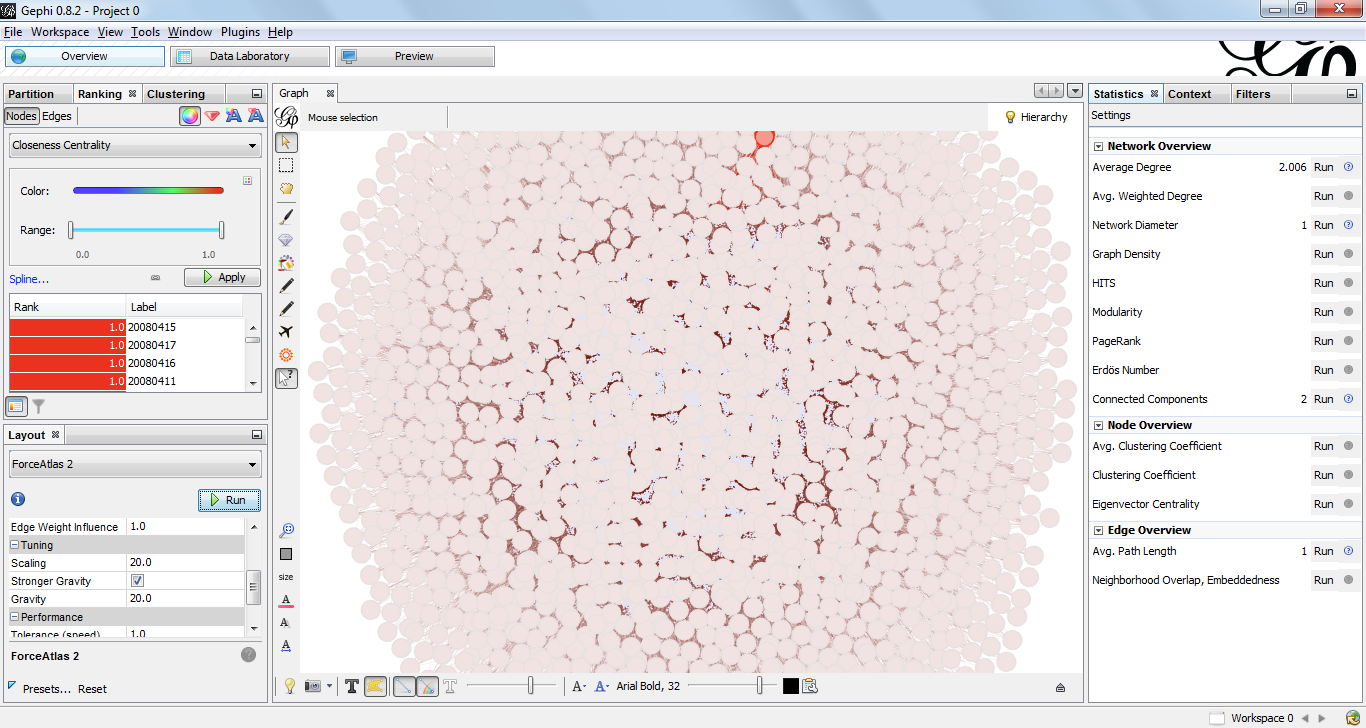
Also we reduce the number of sightings to entries post 2005 (ufo4.graphml) and also post 2010 (ufo5.graphml).



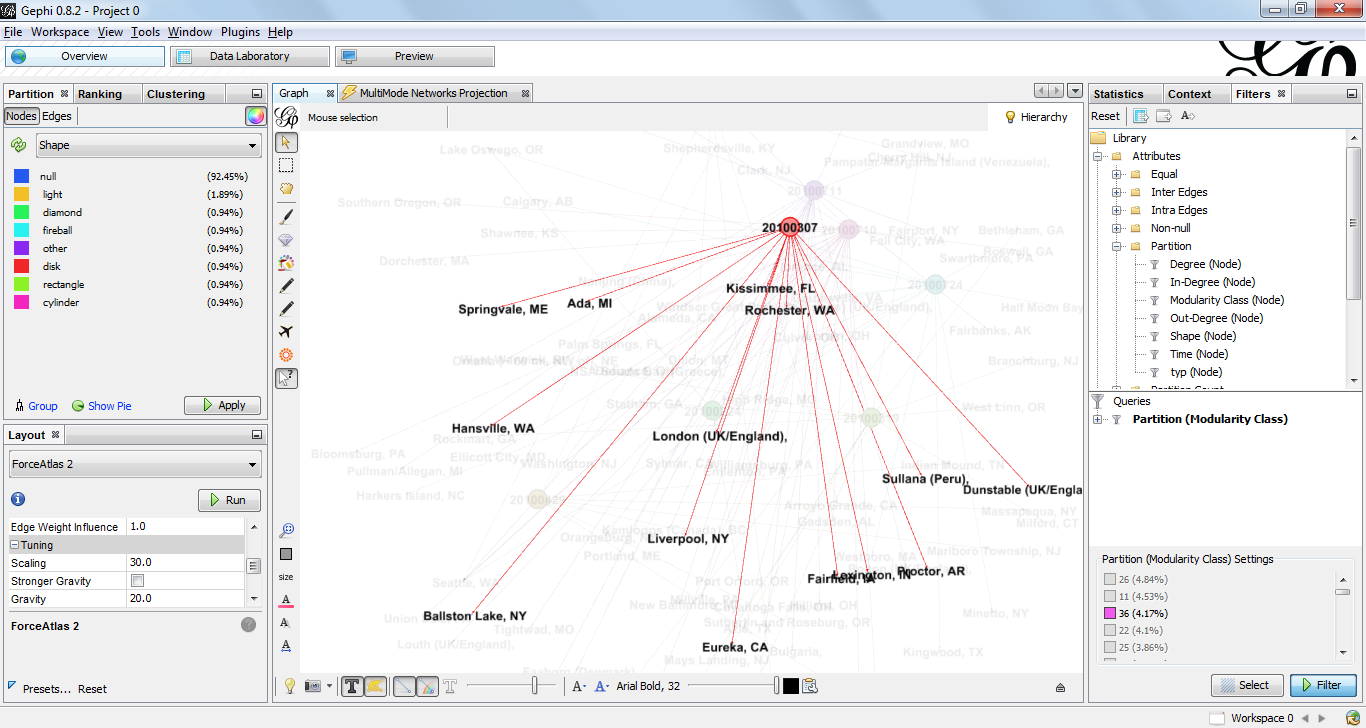


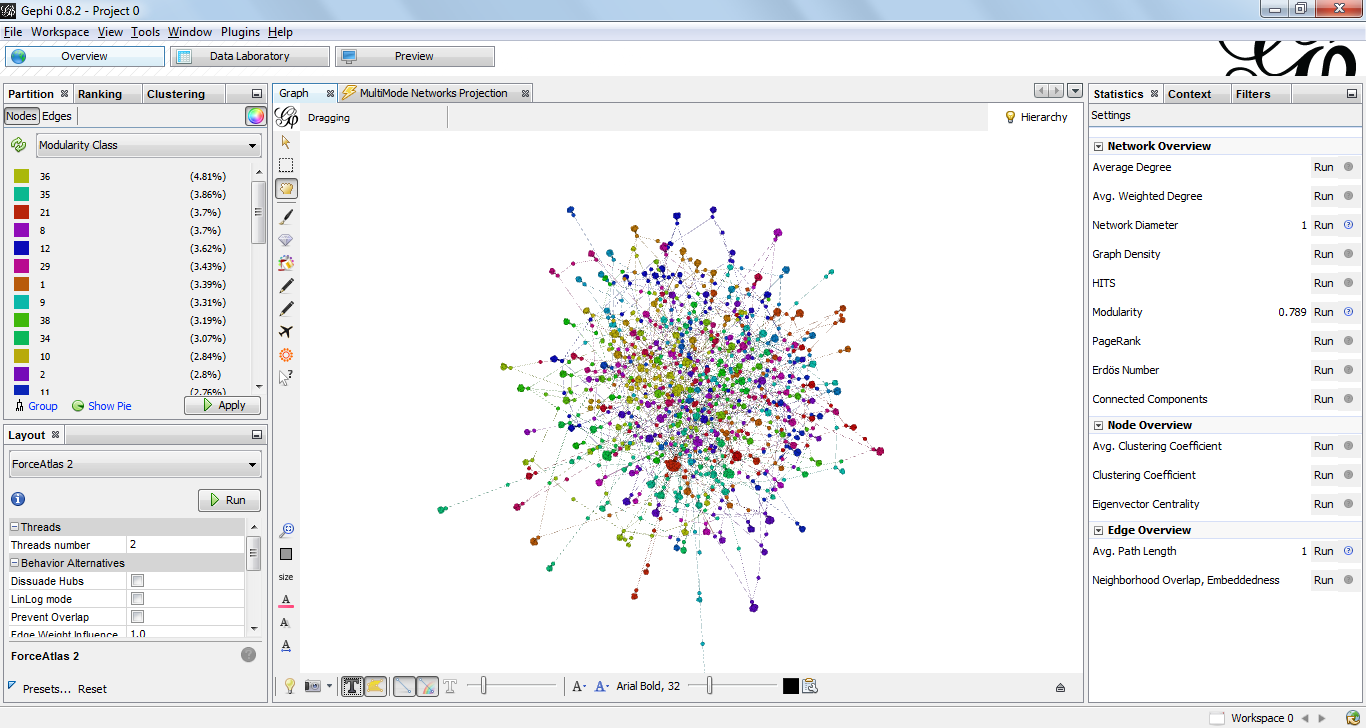
There is no betweenness centrality as there is no relationship between location nodes and Sighting Dates i.e. From data its difficult o ascertain where the UFO’s must have gone post the current location! Its possible to guess that from a location it could be going to next location on same date which could be nearby city, but there is no substantial and consistent data to identify two sightings of same UFO.

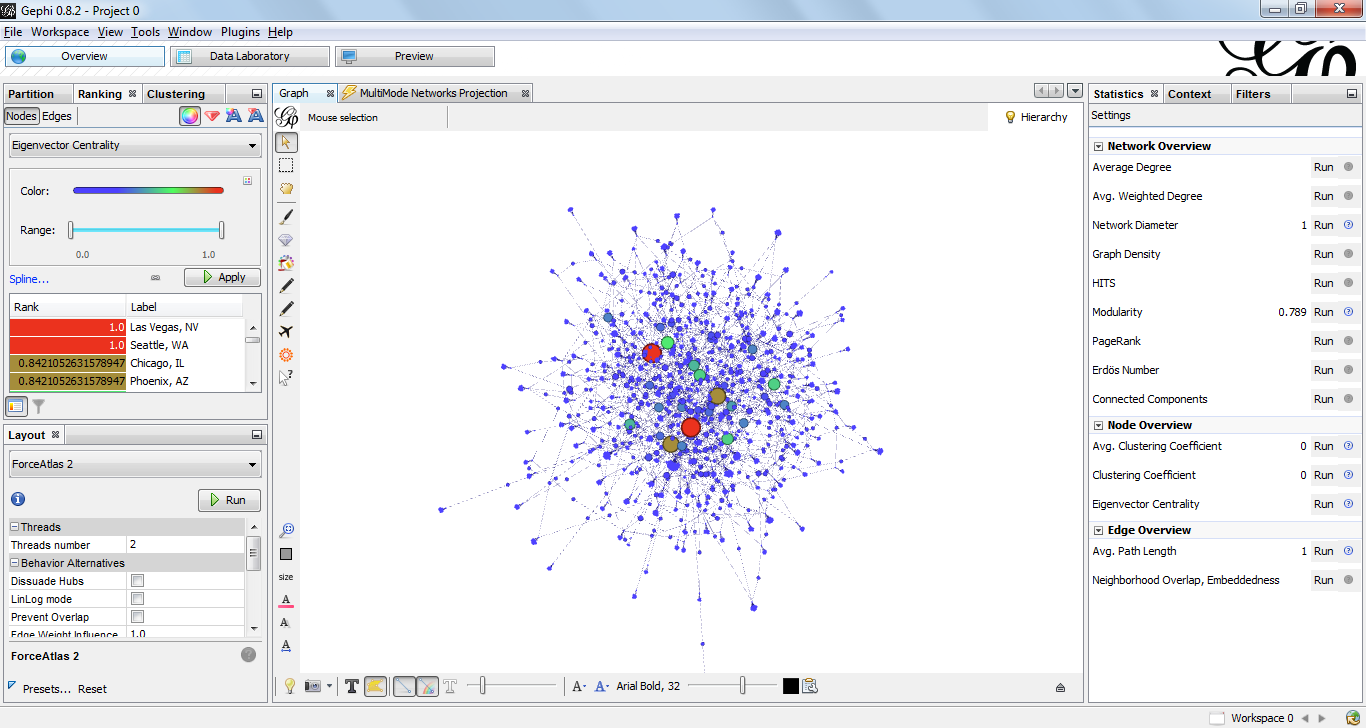
Closeness centrality indicates the location nodes have 0 and Sight Dates have 1.0



Modularity using gephi’s algorithm indicates that there are communities. But it is difficult to pinpoint what can we construed as used to partition the graph.







Eigenvector centrality shows that Las Vegas still highest rank followed by Seattle. I have tried to understand what could be certain issues like linking disasters to sightings but have not been successful. <http://www.ufo-blogger.com/2010/07/disasters-and-ufos.html>

I have examined the disaster databases available here <http://www.em-dat.net/links/disasterdbs.html> but have not found way to link this together. But we have seen much from the data that there may be some pattern in which these UFOs are seen in cities like Phoenix, Las Vegas, Seattle etc. Also it would be interesting to link sightings together to see if these UFO visit different locations as per some design!

# References

1. Slide decks and videos of Professor Lada Adamic
2. <http://ufoevidence.org> for reference on UFO sightings