

SOCIAL NETWORK ANALYSIS

Submitted by :-

Harshil (258 / CO / 14)

Problem

Define the relationship among various groups of people on Twitter with appropriate attributes using NodeXL and visualize clustering among the developed groups.

Introduction

Social network analysis (SNA) is a process of quantitative and qualitative analysis of a social network. SNA measures and maps the flow of relationships and relationship changes between knowledge-possessing entities. Simple and complex entities include websites, computers, animals, humans, groups, organizations and nations.

The SNA structure is made up of node entities, such as humans, and ties, such as relationships. The advent of modern thought and computing facilitated a gradual evolution of the social networking concept in the form of highly complex, graph-based networks with many types of nodes and ties. These networks are the key to procedures and initiatives involving problem solving, administration and operations.

NodeXL is a free and open-source network analysis and visualization software package for Microsoft Excel .It is a set of prebuilt class libraries using a custom Windows Presentation Foundation control. Additional .NET assemblies can be developed as "plug-ins" to import data from outside data providers. Currently-implemented data providers for NodeXL include Facebook, Twitter, Wikipedia (the MediaWiki understructure), web hyperlinks, Microsoft Exchange Server.

Elements

1) Data Import

NodeXL imports UCINet and GraphML files, as well as Excel spreadsheets containing edge lists or adjacency matrices, into NodeXL workbooks. NodeXL also allows for quick collection of social media data via a set of import tools which can collect network data from e-mail, Twitter, YouTube, and Flickr. NodeXL requests the user's permission before collecting any personal data and focuses on the collection of publicly available data, such as Twitter statuses and follows relationships for users who have made their accounts public. These features allow NodeXL users to instantly get working on relevant social media data and integrate aspects of social media data collection and analysis into one tool.

2) Data Representation

NodeXL workbooks contain four worksheets: Edges, Vertices, Groups, and Overall Metrics. The relevant data about entities in the graph and relationships between them are located in the appropriate worksheet in row format. For example, the edges worksheet contains a minimum of two columns, and each row has a minimum of two elements corresponding to the two vertices that make up an edge in the graph. Graph metrics and edge and vertex visual properties appear as additional columns in the respective worksheets. This representation allows the user to leverage the Excel spreadsheet to quickly edit existing node properties and to generate new ones, for instance by applying Excel formulas to existing columns.

3) Graph Analysis

NodeXL contains a library of commonly used graph metrics: centrality, clustering coefficient, diameter. NodeXL differentiates between directed and undirected networks. NodeXL implements a variety of community detection algorithms to allow the user to automatically discover clusters in their social networks.

4) Graph Visualization

NodeXL generates an interactive canvas for visualizing graphs. The project allows users to pick from several well-known Force-directed graph drawing layout algorithms such as Fruchterman-Reingold and Harel-Koren. NodeXL allows the user to multi-select, drag and drop nodes on the canvas and to manually edit their visual properties (size, color, and opacity). In addition, NodeXL allows users to map the visual properties of nodes and edges to metrics it calculates, and in general to any column in the edges and vertices worksheet.

5) Clustering

Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both. If meaningful groups are the goal, then the clusters should capture the natural structure of the data. In some cases, however, cluster analysis is only a useful starting point for other

purposes, such as data summarization. Whether for understanding or utility, cluster analysis has long played an important role in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. There have been many applications of cluster analysis to practical problems. We provide some specific examples, organized by whether the purpose of the clustering is understanding or utility.

Types of Clusters

There are several different notions of cluster that prove useful in practice like :

1) *Well-Separated*

A cluster is a set of objects in which each object is closer (or more similar) to every other object in the cluster than to any object not in the cluster. Sometimes a threshold is used to specify that all the objects in a cluster must be sufficiently close (or similar) to one another. This idealistic definition of a cluster is satisfied only when the data contains natural clusters that are quite far from each other. Well-separated clusters do not need to be globular, but can have any shape.

2) *Prototype-Based*

A cluster is a set of objects in which each object is closer (more similar) to the prototype that defines the cluster than to the prototype of any other cluster. For data with continuous attributes, the prototype of a cluster is often a centroid, i.e., the average (mean) of all the points in the cluster. When a centroid is not meaningful, such as when the data has categorical attributes, the prototype is often a medoid, i.e., the most representative point of a cluster.

3) *Graph-Based*

If the data is represented as a graph, where the nodes are objects and the links represent connections among objects, then a cluster can be defined as a connected component i.e., a group of objects that are connected to one another, but that have no connection to objects outside the group. An important example of graph-based clusters are contiguity-based clusters, where two objects are connected only if they are within a specified distance of each other. This implies that each object in a contiguity-based cluster is closer to some other object in the cluster than to any point in a different cluster.

Advantage of this Project

This project can be used to visualize the various relationships among people in a real world scenario .The task of collecting user database manually is a very cumbersome and time inefficient task. Instead this project can be used to fetch data from popular networking sites with a few clicks. Also the simplistic and minimalistic GUI of the project will make it easier for anyone to work with SNA without much hassle. This project can fetch large amount of data , provide choice between clustering algorithm and also show the metrics and miscellaneous data obtained after analysing.

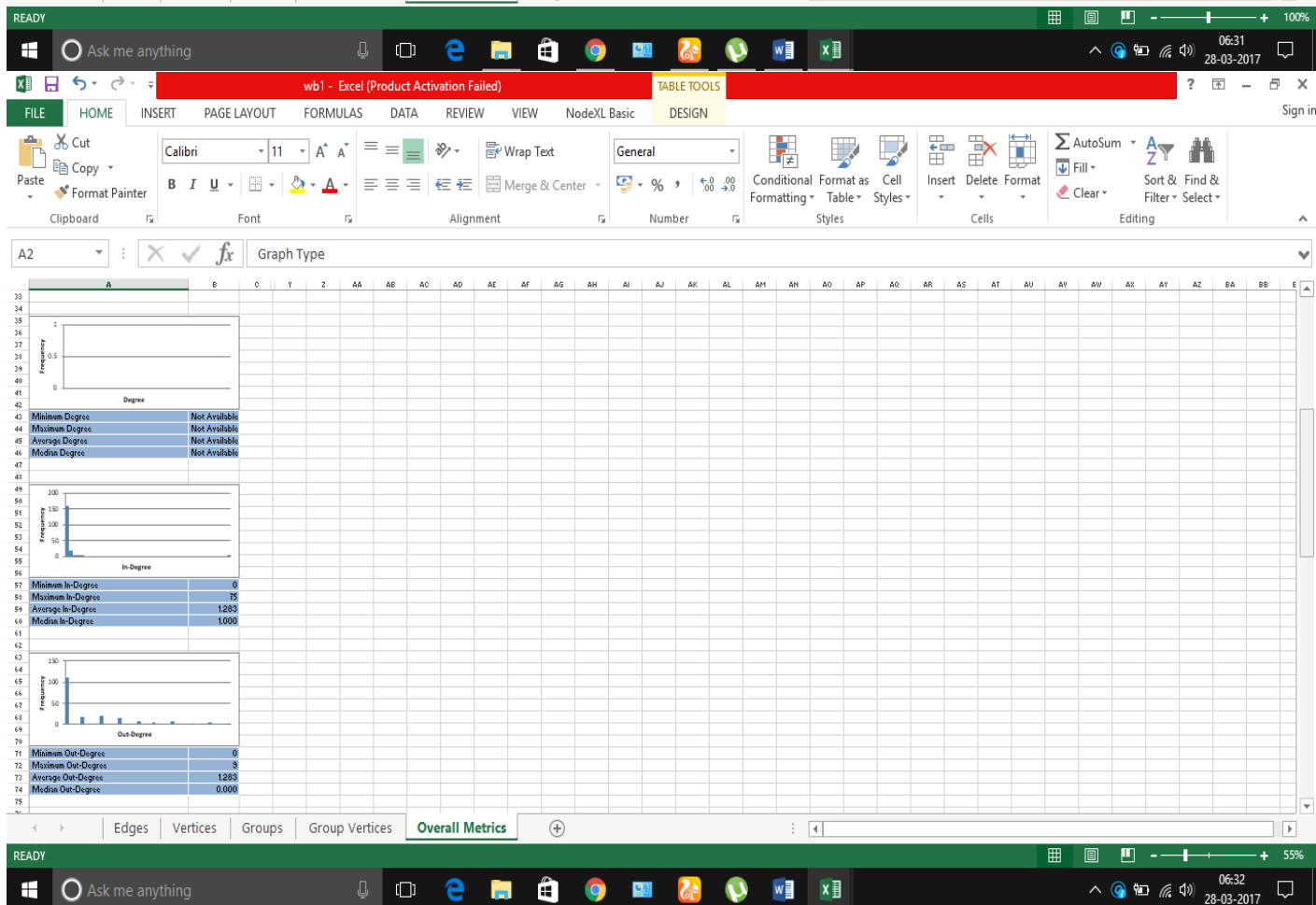
Implementation

- 1)Data Imported from Twitter for a particular keyword in NodeXL.
- 2)Cluster the data using appropriate attributes and algorithm.
- 3)Analyze the social network from the graph.
- 4)Obtain the graph metrics for detailed and in-depth information.

Excel interface showing a spreadsheet with columns for Visual Properties, Labels, Graph Metrics, and Relationships. The spreadsheet is titled "wb1 - Excel (Product Activation Failed)". The interface includes a ribbon with tabs like FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEW, VIEW, NodeXL Basic, and DESIGN. The spreadsheet data includes columns for Vertex, Color, Width, Style, Opacity, Visibility, Label, Color, Size, Reciprocity, Add Your Own Columns Here, Relationship, Date (UTC), Tweet, URL, Domains, HashTags, Tweet Date (UTC), Twitter Page, Latitude, Longitude, Import, and In-Reply. The spreadsheet is displaying data for a network graph, with vertices and edges represented by colored cells. A tooltip for "Vertex 1 Name" is visible, indicating the name of the edge's first vertex.

Excel interface showing a spreadsheet with columns for Visual Properties, Labels, Graph Metrics, and Relationships. The spreadsheet is titled "wb2 - Excel (Product Activation Failed)". The interface includes a ribbon with tabs like FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEW, VIEW, NodeXL Basic, and DESIGN. The spreadsheet data includes columns for Vertex, Color, Width, Style, Opacity, Visibility, Label, Color, Size, Reciprocity, Add Your Own Columns Here, Relationship, Date (UTC), Tweet, URL, Domains, HashTags, Tweet Date (UTC), Twitter Page, Latitude, Longitude, Import, and In-Reply. The spreadsheet is displaying data for a network graph, with vertices and edges represented by colored cells. A tooltip for "Vertex 1 Name" is visible, indicating the name of the edge's first vertex.

Graph Metric		Value											
Graph Type		Directed											
Vertices		187											
Unique Edges		221											
Edges With Duplicates		45											
Total Edges		266											
Self-Loops		0											
Reciprocated Vertex Pair Ratio		0.0041841											
Reciprocated Edge Ratio		0.008333333											
Connected Components		2											
Single-Vertex Connected Components		0											
Maximum Vertices in a Connected Component		178											
Maximum Edges in a Connected Component		258											
Maximum Geodesic Distance (Diameter)		4											
Average Geodesic Distance		3.067275											
Graph Density		0.006900121											



wb2 - Excel (Product Activation Failed)

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW NodeXL Basic DESIGN

Normal Page Break Preview Custom Views Workbook Views Ruler Formula Bar Gridlines Headings Document Actions Zoom 100% Zoom to Selection New Window Arrange All Freeze Panes Hide Synchronous Scrolling Reset Window Position Switch Windows Macros

A9

Frequency

In-Degree

Minimum In-Degree 0

Maximum In-Degree 39

Average In-Degree 1.697

Median In-Degree 0.000

Frequency

Out-Degree

Minimum Out-Degree 0

Maximum Out-Degree 19

Average Out-Degree 1.697

Median Out-Degree 1.000

Edges Sheet1 Vertices Groups Group Vertices Overall Metrics

READY

Ask me anything

wb1 - Excel (Product Activation Failed)

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW NodeXL Basic DESIGN

Import Export Prepare Data Refresh Graph Summary Automate Type: Directed Layout: Harel-Koren... Color Vertex Shape Vertex Size Opacity Visibility Edge Width Dynamic Filters Graph Metrics Subgraph Images Groups Use Current for New Import Export Reset All Workbook Columns Graph Elements Show Notifications Online Upgrade About Help

Data Graph Visual Properties Analysis Options Show/Hide Help

A2

Graph Type

Graph Metric Value

Graph Type Directed

Vertices 187

Unique Edges 221

Edges With Duplicates 45

Total Edges 266

Self-Loops 0

Reciprocal Vertex Pair Ratio 0.0041941

Reciprocal Edge Ratio 0.0083333

Connected Components 2

Single-Vertex Connected Components 0

Maximum Vertices in a Connected Component 118

Maximum Edges in a Connected Component 258

Maximum Geodesic Distance (Diameter) 4

Average Geodesic Distance 3.067215

Graph Density 0.0065001

Modularity Not Applicable

NodeXL Version 1.0.1361

Readability Metric Value

Frequency

Degree

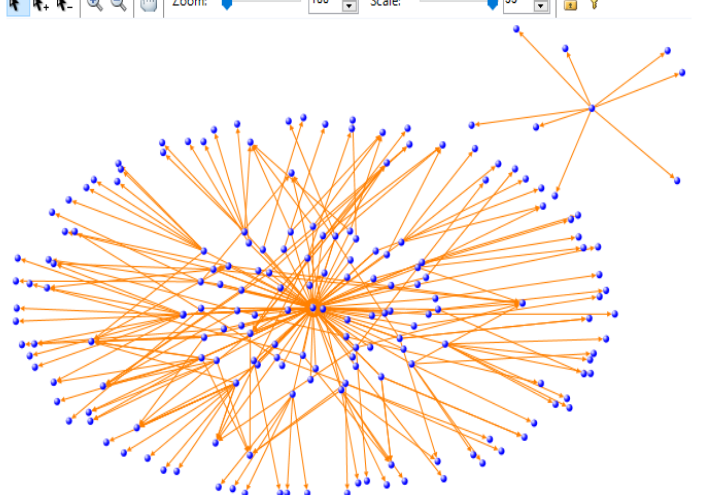
Minimum Degree Not Available

Overall Metrics

Document Actions

Refresh Graph Harel-Koren Fast Mul Lay Out Again Dynamic Filters Graph Options

Zoom: 100 Scale: 99

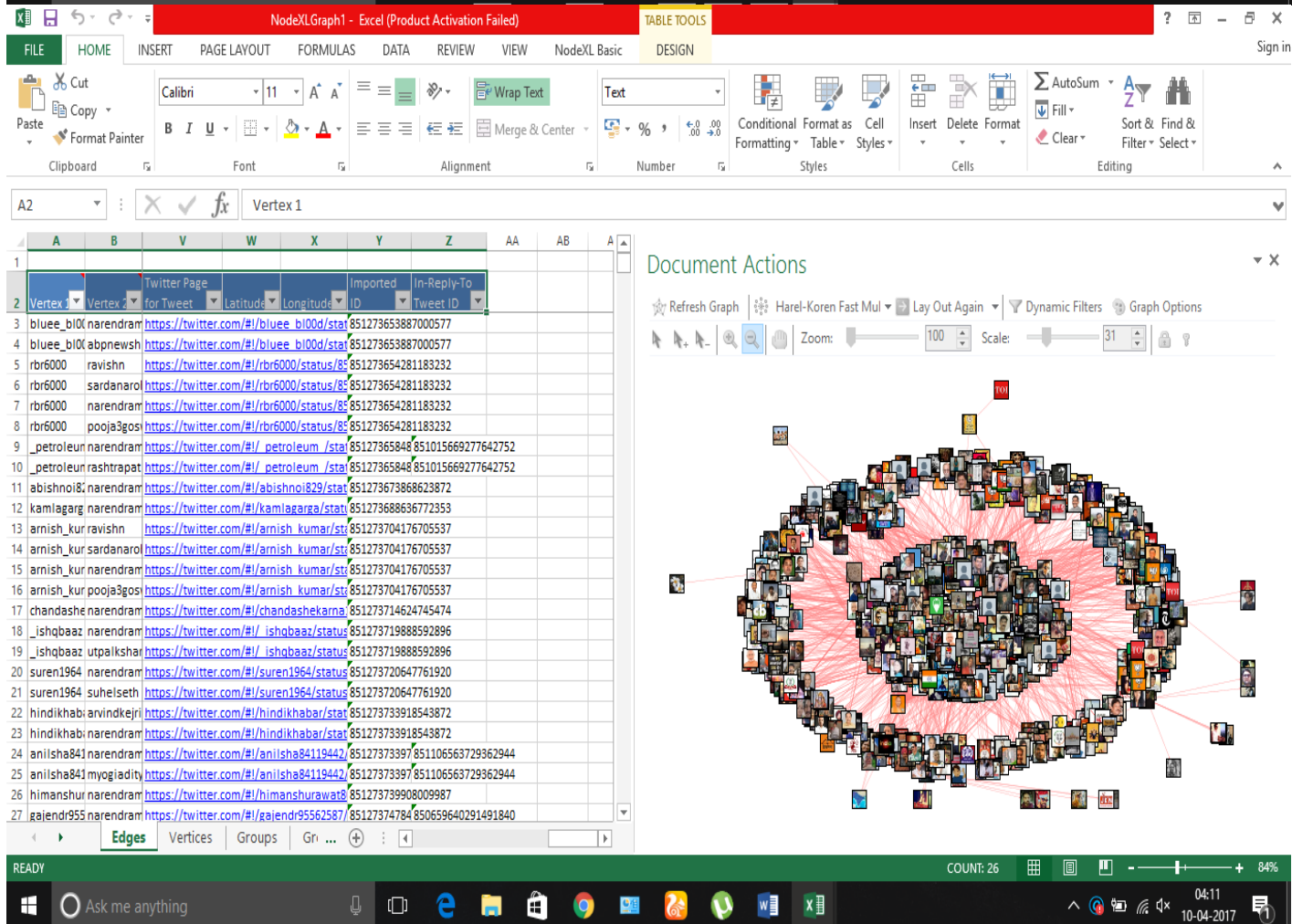
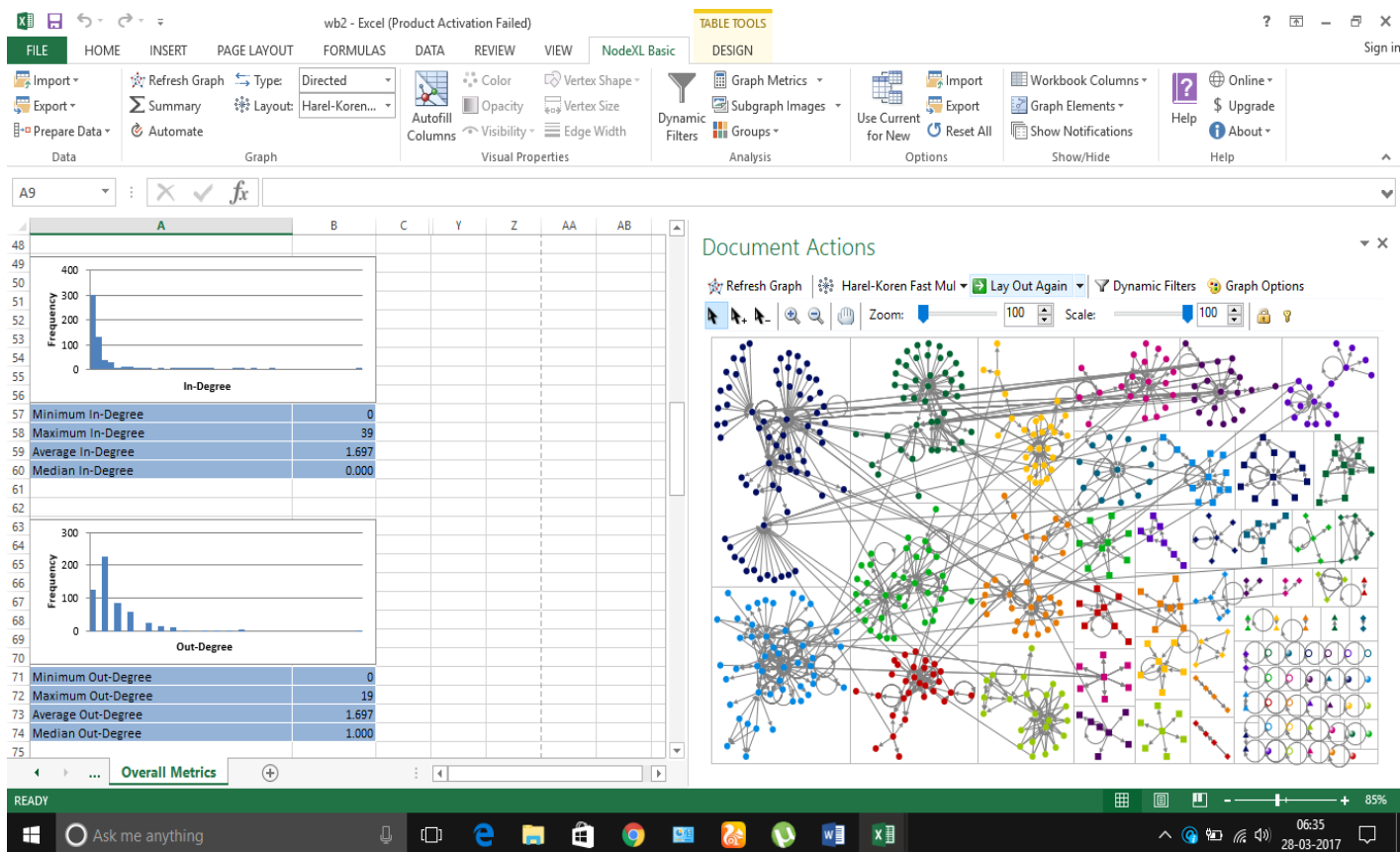


READY

Ask me anything

55%

06:34 28-03-2017



wb1 - Excel (Product Activation Failed) | TABLE TOOLS | DESIGN

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW NodeXL Basic

Clipboard Font Alignment Number Styles Cells Editing

Calibri 11 A A Wrap Text B I U Merge & Center Text % .00 .00 Conditional Formatting Table Cell Styles Insert Delete Format AutoSum Fill Sort & Find & Filter Select

A266 : arvindt38484730

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Visual Properties													
2	Vertex	Vertex	Color	Width	Style	Opacity	Visibility	Label	Label Text	Label Color	Label Font Size	Graph Metrics	Other Columns	
251	shubhami	narendramodi												
252	shoksing	nationality/L												
253	shoksing	omoficeup												
254	shoksing	pmoindia												
255	shoksing	situjayal												
256	shoksing	narendramodi												
257	shoksing	narendramodi												
258	harenderb	thekernelspeaks												
259	harenderb	yashkin5												
260	harenderb	sanghaviacepa												
261	harenderb	saundid												
262	harenderb	narendramodi												
263	saniavsup	narendramodi												
264	arvindt384	number10gov												
265	arvindt384	theresa_may												
266	arvindt384	narendramodi												
267	kullas													
268	kullas													
269														
270														
271														
272														
273														
274														
275														
276														
277														
278														
279														
280														
281														

Vertex 1 Name
Enter the name of the edge's first vertex.

Document Actions

Refresh Graph Harel-Koren Fast Mul Lay Out Again Dynamic Filters Graph Options

Zoom: 100 Scale: 99

Edges Vertices Groups Gr ...

READY AVERAGE: 28592.26226 COUNT: 909 SUM: 8063017.959 69%

Ask me anything 02:11 10-04-2017

wb1 - Excel (Product Activation Failed) | TABLE TOOLS | DESIGN

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW NodeXL Basic

Import Export Prepare Data Refresh Graph Summary Automate Type: Directed Layout: Fruchterman... Color Vertex Shape Opacity Visibility Autofill Columns Dynamic Filters Graph Metrics Subgraph Images Groups Use Current for New Import Export Workbook Columns Graph Elements Show Notifications Help Upgrade About

A2 : Group

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Visual Properties													
2	Group	Vertex Color	Vertex Shape	Visibility	Collapsed?	Label	Vertices	Edges						
3	G1	0, 12, 96	Disk											
4	G2	0, 136, 227	Disk											
5	G3	0, 100, 50	Disk											
6	G4	0, 176, 22	Disk											
7	G5	191, 0, 0	Disk											
8	G6	230, 120, 0	Disk											
9	G7	255, 191, 0	Disk											
10	G8	150, 200, 0	Disk											
11	G9	200, 0, 120	Disk											
12	G10	77, 0, 96	Disk											
13	G11	91, 0, 191	Disk											
14	G12	0, 98, 130	Disk											
15	G13	0, 12, 96	Solid Square											
16	G14	0, 136, 227	Solid Square											
17	G15	0, 100, 50	Solid Square											
18	G16	0, 176, 22	Solid Square											
19														
20														
21														
22														
23														

Document Actions

Refresh Graph Fruchterman-Reingo Lay Out Again Dynamic Filters Graph Options

Zoom: 100 Scale: 99

Vertices Groups Group Vertic ...

READY COUNT: 24 100%

Ask me anything 02:15 10-04-2017

wb1 - Excel (Product Activation Failed)

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW NodeXL Basic DESIGN

Import Export Prepare Data Refresh Graph Summary Automate Type: Directed Layout: Fruchterman... Color Opacity Visibility Autofill Columns Vertex Shape Vertex Size Edge Width Dynamic Filters Graph Metrics Subgraph Images Groups Use Current for New Import Export Reset All Workbook Columns Graph Elements Show Notifications Online Upgrade About Help

Data Graph Visual Properties Analysis Options Show/Hide Help

A266 arvindt38484730

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2	Vertex	Vertex	Color	Width	Style	Opacity	Visibility	Label	Label Text	Label Font Color	Label Font Size	Reciprocity		
251	shubhami	narendramodi												
252	shoksing	nationalityL												
253	shoksing	omoficeup												
254	shoksing	pmoindia												
255	shoksing	situjayal												
256	shoksing	narendramodi												
257	shoksing	narendramodi												
258	harenderb	thekernalspeaks												
259	harenderb	yashkir5												
260	harenderb	sanghviceepa												
261	harenderb	saundid												
262	harenderb	narendramodi												
263	saniaysup	narendramodi												
264	arvindt384	number10gov												
265	arvindt384	theresa_may												
266	arvindt384	narendramodi												
267	kulla													
268	kulla													
269														
270														
271														
272														
273														
274														
275														
276														
277														
278														
279														
280														
281														

Vertex 1 Name
Enter the name of the edge's first vertex.

Edges Vertices Groups Gri ...

Document Actions

Refresh Graph Fruchterman-Reingo Lay Out Again Dynamic Filters Graph Options

Zoom: 100 Scale: 99

AVERAGE: 28596.06698 COUNT: 262 SUM: 2144705.023

02:19 10-04-2017

wb1 - Excel (Product Activation Failed)

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW NodeXL Basic DESIGN

Import Export Prepare Data Refresh Graph Summary Automate Type: Directed Layout: Fruchterman... Color Opacity Visibility Autofill Columns Vertex Shape Vertex Size Edge Width Dynamic Filters Graph Metrics Subgraph Images Groups Use Current for New Import Export Reset All Workbook Columns Graph Elements Show Notifications Online Upgrade About Help

Data Graph Visual Properties Analysis Options Show/Hide Help

A266 arvindt38484730

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2	Vertex	Vertex	Color	Width	Style	Opacity	Visibility	Label	Label Text	Label Font Color	Label Font Size	Reciprocity		
251	shubhami	narendramodi												
252	shoksing	nationalityL												
253	shoksing	omoficeup												
254	shoksing	pmoindia												
255	shoksing	situjayal												
256	shoksing	narendramodi												
257	shoksing	narendramodi												
258	harenderb	thekernalspeaks												
259	harenderb	yashkir5												
260	harenderb	sanghviceepa												
261	harenderb	saundid												
262	harenderb	narendramodi												
263	saniaysup	narendramodi												
264	arvindt384	number10gov												
265	arvindt384	theresa_may												
266	arvindt384	narendramodi												
267	kulla													
268	kulla													
269														
270														
271														
272														
273														
274														
275														
276														
277														
278														
279														
280														
281														

Vertex 1 Name
Enter the name of the edge's first vertex.

Edges Vertices Groups Gri ...

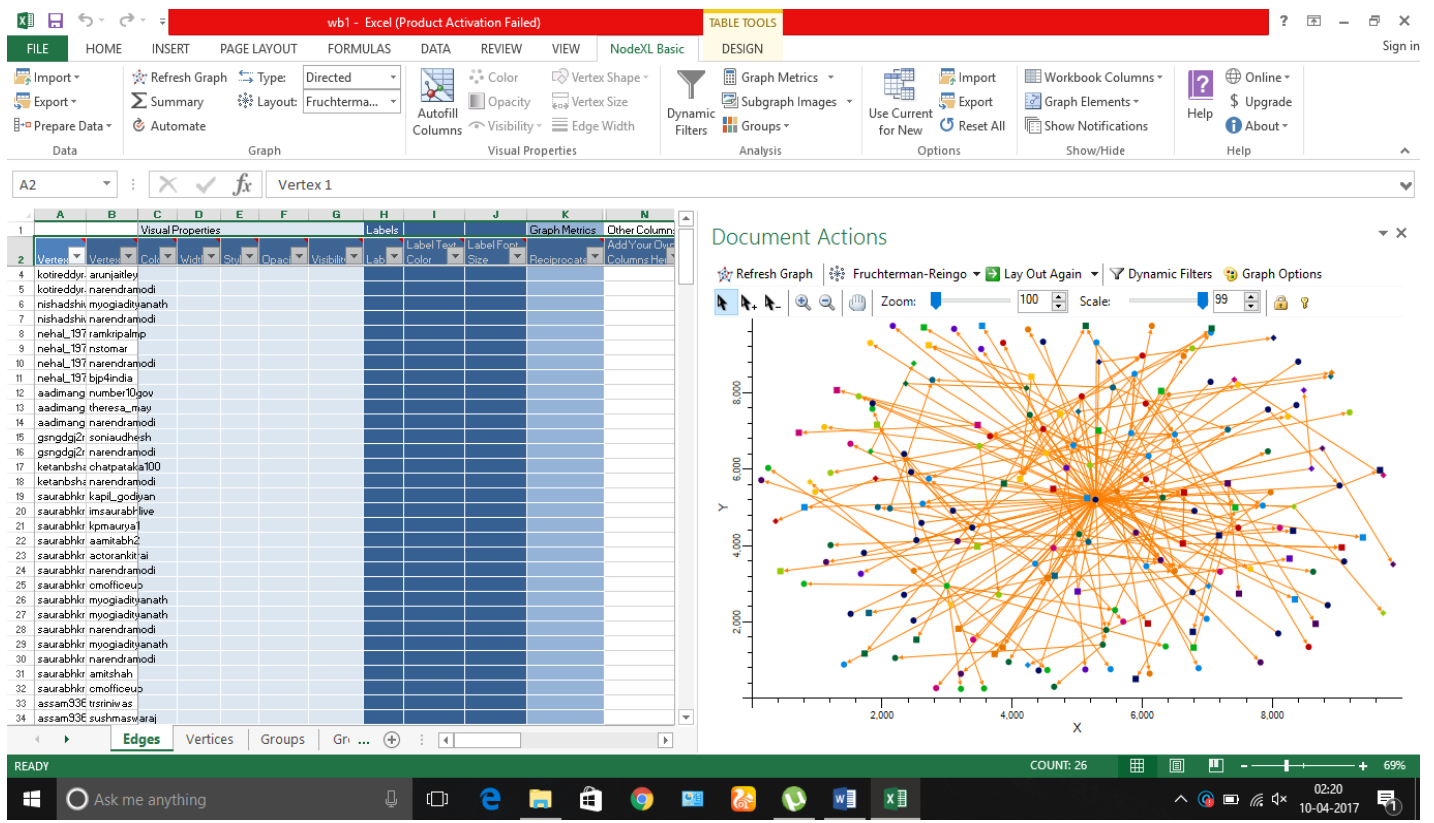
Document Actions

Refresh Graph Fruchterman-Reingo Lay Out Again Dynamic Filters Graph Options

Zoom: 100 Scale: 99

AVERAGE: 28592.26226 COUNT: 1003 SUM: 8063017.959

02:20 10-04-2017



Algorithms Used

1) Clauset-Newman-Moore Algorithm

This algorithm is useful for large input data and lots of relations involved. This divisive algorithm that uses edge difference as a metric to identify the boundaries of communities. This algorithm has been applied successfully to a variety of networks, including networks of email messages, human and animal social networks, networks of collaborations between scientists and musicians, metabolic networks and gene networks. The algorithm proposed uses a greedy optimization in which, starting with each vertex being the sole member of a community of one, we repeatedly join together the two communities whose amalgamation produces the largest increase.

We maintain three data structures :

1. A sparse matrix containing ΔQ_{ij} for each pair i, j of communities with at least one edge between them. We store each row of the matrix both as a balanced binary tree

(so that elements can be found or inserted in $O(\log n)$ time) and as a maxheap (so that the largest element can be found in constant time).

2. A max-heap H containing the largest element of each row of the matrix ΔQ_{ij} along with the labels i, j of the corresponding pair of communities.
3. An ordinary vector array with elements a_i .

Algorithm can now be defined as follows.

1. Calculate the initial values of ΔQ_{ij} and a_i , and populate the max-heap with the largest element of each row of the matrix ΔQ .
2. Select the largest ΔQ_{ij} from H , join the corresponding communities, update the matrix ΔQ , the heap H and a_i (as described below) and increment Q by ΔQ_{ij} .
3. Repeat step 2 until only one community remains.

2) Girvan Newman Algorithm

The Girvan–Newman algorithm (named after Michelle Girvan and Mark Newman) is a hierarchical method used to detect communities in complex systems. It detects communities by progressively removing edges from the original network. The connected components of the remaining network are the communities. Instead of trying to construct a measure that tells us which edges are the most central to communities, the Girvan–Newman algorithm focuses on edges that are most likely "between" communities. The algorithm's steps for community detection are summarized below :

1. The betweenness of all existing edges in the network is calculated first.
2. The edge with the highest betweenness is removed.
3. The betweenness of all edges affected by the removal is recalculated.
4. Steps 2 and 3 are repeated until no edges remain.

By recalculating betweennesses after the removal of each edge, it is ensured that at least one of the remaining edges between two communities will always have a high value. The end result of the Girvan–Newman algorithm is a dendrogram. As the Girvan–Newman algorithm runs, the dendrogram is produced from the top down (i.e. the network splits up into different communities with the successive removal of links). The leaves of the dendrogram are individual nodes.

Applications of this Project

- 1) Industries : finding how well a group of people dwell with co- workers and make optimal use of resources for more profits.*
- 2) Public Sector : that uses development of leader engagement strategies, analysis of individual and group engagement and media use, and community-based problem solving.*
- 3) Marketing : finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records.*
- 4) Insurance : identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds.*
- 5) City-Layout and Designing : identifying groups of houses according to their house type, value and geographical location.*

Bibliography

- 1. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/*
- 2. https://en.wikipedia.org/wiki/Social_network_analysis*
- 3. http://www.uvk.de/uploads/tx_gbuvkbooks/PDF_L/9783867640466_L.pdf*
- 4. https://paolatubaro.files.wordpress.com/2012/03/introductionchair_v1.pdf*