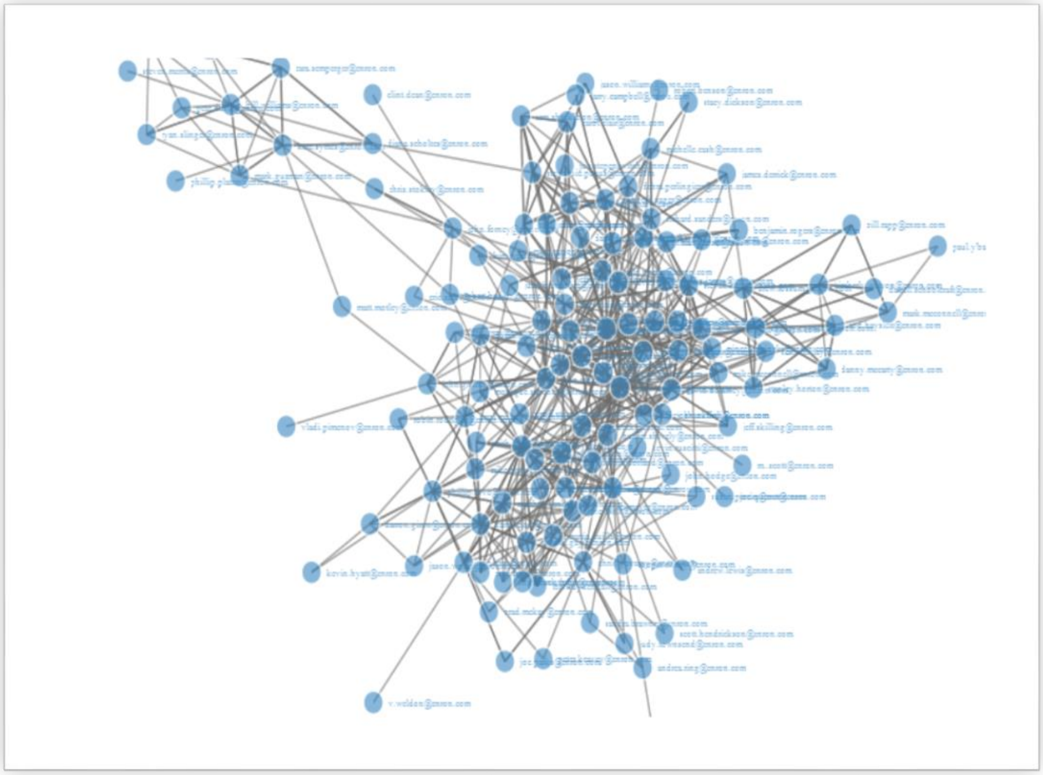


Enron Dataset Analysis using Graph Algorithms

FINDING MEASURES OF CENTRALITY IN A SOCIAL NETWORK

Rachita Bansal | Program Structure & Algorithms | August 20, 2015

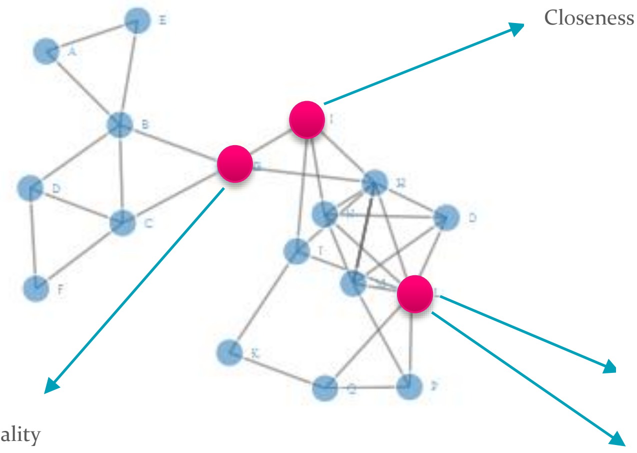
Introduction

Centrality

Centrality refers to relative measure of importance of a node in the graph. Centrality measures are important to analyze and find how influential a person is in social networks.

Measures of Centrality

1. Degree Centrality
2. Closeness Centrality
3. Betweenness Centrality
4. Eigenvector Centrality



Closeness Centrality

Betweenness Centrality

Eigenvector Centrality

Degree Centrality

Degree Centrality

It's the most intuitive notion of centrality. Node with the highest degree is the most important. It gives the index of exposure to what is flowing in the network. Central actor is most likely to hear a gossip.

In a graph, it's given by:-

$$\text{centrality} \;=\; (u)\text{i} \;=\sum_{ij}^{n}(\text{X})j$$

In a directional graph, it's given by measure of out-degree and in-degree:-

Out-degree centrality = ∑out bounds from a node
In-degree centrality = ∑in-bounds to a node

Betweenness Centrality

Betweenness centrality of a node 'u' is the ratio of the shortest paths between all other nodes that pass through u.

In a graph, it's given by:-

$$CB\ (u)\ =\sum_{s\neq v\neq t}(\mathrm{\sigma st(u)/\sigma st})$$

CB (u) = Betweenness centrality of node u
σst (u) = number of shortest paths between (s, t) that passes through u
σst = number of shortest paths between (s, t)

Closeness Centrality

A node is considered important if it is relatively close to all other nodes. It's a measure of how long it will take to spread information from node 'u' to all other nodes.

Farness is the sum of a node's distances to all other nodes.

Closeness is the inverse of Farness. For disconnected networks, It's given by:-

$$CC\ (u)\ = 1/\sum_{v\neq u}(\mathrm{\sigma(u,v)})$$

Reference: http://toreopsahl.com/2010/03/20/closeness-centrality-in-networks-with-disconnected-components/

Eigenvector Centrality

It's the measure of influence of a node in a network. Connections to high scoring nodes contribute more. "An important node is connected to important neighbor" which means a node has high score if connected to many nodes are themselves well connected. Power iteration is one of the eigenvector algorithm.

Centralization of Network

Measure of how central its most central node is in relation to how central all other nodes are. Measures the extent to which network revolves around a single node.

It's given by:-

$$CD\ (u)\ =\sum_{i=1}^{n}\left(\left|\frac{\mathrm{Cd(n)}-Cd(i)}{\mathrm{(N-1)(N-2)}}\right|\right)$$

Enron Email Dataset

Enron was founded in 1985. It started as energy business and gas firm which later expanded to other projects becoming the 7th largest business organization in the USA over 15 years. Enron declared bankruptcy in December of 2001 which was followed by several investigations. During the investigations Enron Email Dataset consisting of around 6000,000 emails was made public on the web. The dataset was purchased by Lesie Kaelbling at MIT and was later posted by Prof. William W. Cohen at CMU.

Organization of the Dataset

The current version has 517,431 mails organized into 150 folders. The folder name is given as employee last name followed by a hyphen and first letter of the first name. For example, "allen-p" is named after Phillip K. Allen. Each employee folder consists of multiple subfolders, such as, "inbox", "deleted_items", "_sent_mail", "discussion_threads" and others created by the employee. The data is dated from its glory to collapse, Nov. 2008 to June 2002.

Dataset was downloaded from:

https://www.cs.cmu.edu/~./enron/

Analysis Methodology

The first step consist of finding the centrality measures like degree, closeness, betweenness and eigenvector centrality to find who has received most emails and who is the most active person. Which pairs are communicating most frequently, this information is important to understand communication patterns. This can also be helpful to cluster people into cohesive subgroups in which people talk to each other more intensively than with outsiders.

In the content analysis, the word count method is used to count the frequency of a bag of keywords in each email. Since these emails are sent by a group of people, the frequency of the bag of keywords can be arrogated on individual level. As a result, each individual has a pattern of word usage. It is interesting to detect relationship between people's characteristics and their communication frequency. For example, if two people having similar usage of words do they communicate frequently.

Properties of Email Data: Data Extracted for Analysis

An Internet email message is divided into two parts the header and body. The header contains structured data namely, From, To, Subject, date and Time. The body contains unstructured data including the content and sometimes a signature.

This information was extracted to establish who talks to who and build a network of all the people who communicate with each other keeping the information such as the sender, receiver, Cc, bcc, subject and first 250 characters of the body. The rest of the information has been ignored or not used in this analysis.

Cleaning Email Dataset: Procedure and Problems faced

Like other raw forms of data, the Enron email dataset is noisy and needed to be cleaned. In general the email data had three problems,

1. Duplicate email addresses exist in the dataset. These are aliases, labeling the same person. Moreover one person can have various emails from domain in addition to his or her organizational one for example, yahoo, Hotmail etc. When people of who send and receive emails are of interest in SNA, mislabeling might lead to incorrect results.
2. Duplicate email messages exist. For example if A sends an email to B, the mail would be in 'outbox' of A and 'Inbox' of B. Also, if there are multiple recipients, all of them will have a copy of the email. Duplicate Emails must be removed or word frequency is of concern, also the frequency of communication between the sender and recipients is of concern as well.
3. Content of the email is difficult to extract. The email content is generally mingled with signature and special characters in the header which take unnecessary space and not a part of the analysis. For example, hyphens in the subject headers.

Procedure followed to identify aliases in the dataset:

Single format was considered to identify the primary email address of a person, which has the first name followed by a "." And then last.

References: http://search.proquest.com/docview/304506228

The program was run on the raw data consisting of 517,431 emails and all unique email addresses in the "From" and "To" part of the message header were extracted. A properties file called "duplicates" was created which consists of the mapping of all aliases of a person to his or her primary email. For example, Phillip.K.Allen@enron.com and pallen@enron.com are mapped to phillip.enron@enron.com.

The file keeps the mapping which has a, "#" sign preceding the name of the person at the top and its corresponding email mappings in the following lines one followed by the other. The aliases are separated by the primary email address from "=" sign. For example, mapping for Phillip Allen would be,

```
#Phillip Allen
Phillip.K.Allen@enron.com=phillip.allen@enron.com
pallen@enron.com=phillip.allen@enron.com
```

Uninteresting Message Identification

On further analysis, it was found that more than 500 email groups/distribution lists such no.address@enron.com, messages@enron.com, bills@enron.com, announcements@enron.com, and so on, exist in the dataset which have been removed from the analysis. As these lists are assumed to send generic posts circulated in the organization they happen to have large out-degree and very less in-degree such as 0, 1, 2, which was a concern in the analysis. The role of these lists has been ignored in the data analysis.

A properties file called "blacklist" was created to cater to this problem. This list removes all the emails that are sent by the groups that are mentioned in list of addresses specified in this file.

Procedure followed to remove duplicate messages in the dataset:

The mail consists of header which has the sender, receiver subject and the body having the content of the mail. These four factors were considered to identify a duplicate email. If the emails have all of these parameters as same are considered to be same mails and are knocked off from the analysis. This prevents duplication of emails in the dataset.

Also, by manual research over the internet it was found that subfolders in the dataset called “all_documents”, “notes_inbox” and “discussion_Threads” were computer generated folders during the transformation of data and contained a large amount of duplicate emails. Hence, they are moved out from this analysis and around 200,000 emails remain for analysis.

Procedure to extract the content of emails

Due to the memory limitation and huge dataset only the first 250 characters of the content were extracted from the body. The unnecessary hyphens were removed from the subject line so that maximum message is retained.

Data Analysis using Social Network Analysis Methods

Centrality Analysis

A series of centrality measures that describe the ‘importance’ of a person are computed and reviewed. Centrality measures how much a node is involved in the network. Involvement includes both sending and receiving mails. Adjacency matrix was used as a graph notation to represent relationship between the nodes which a *n* x *n* matrix and each cell is a 0 or 1, in which, 1 represent a link from node in the row to the node in the column.

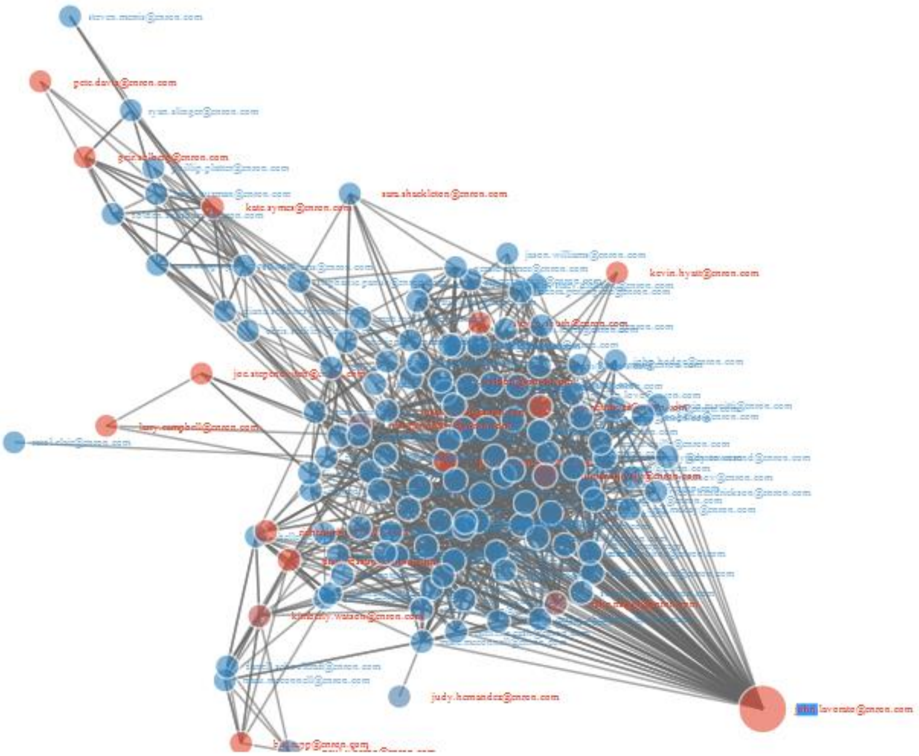
Degree Centrality

Degree of a node in a directional relationship is the In-degree and the Out-degree incident to and from itself to other nodes respectively. Freeman approach was used to determine the degree which considers which calculates links to and from form a node.

A node with high degree centrality signifies more visibility in the network, which is also recognized as an active member and a major channel of information in the network. Since they have many connections, they have many choices in satisfying their needs and hence are less dependent on others.

Top 10 nodes with most In-degree and Out-degree

ID	34	33	73	16	76	72	65	35	32	50
InDeg	45	41	36	35	34	31	31	30	28	27
ID	34	74	33	68	16	58	76	30	114	69
OutDeg	78	71	41	38	37	35	33	33	12	30



Degree Centrality in the Enron dataset

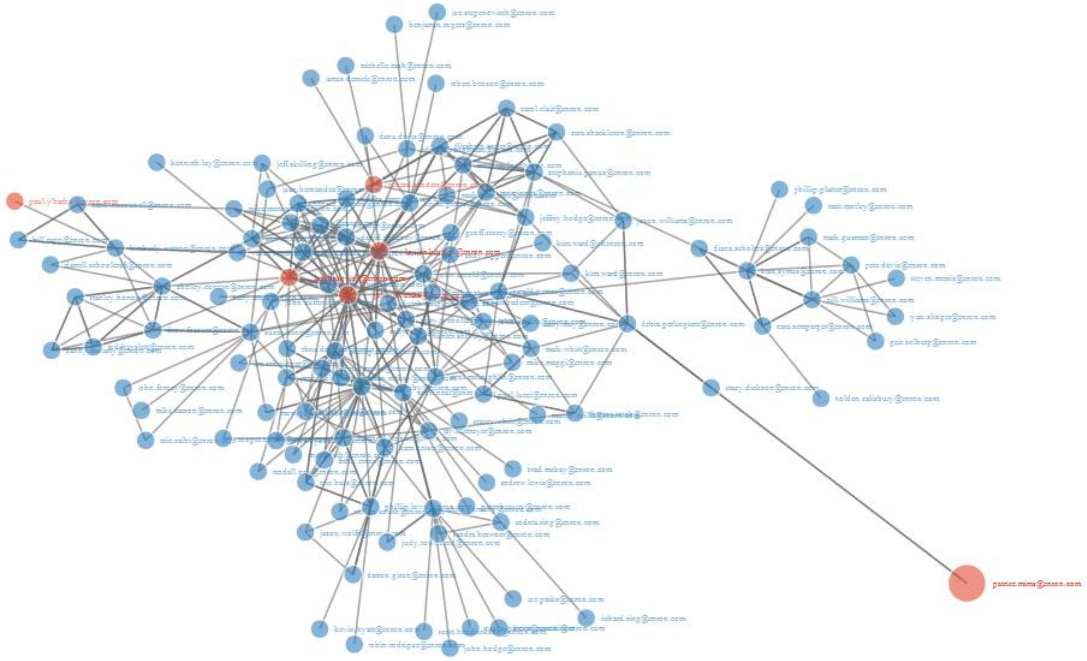
Closeness Centrality

Closeness centrality measures distance of a node to all other nodes in the network. The closer the node is to other nodes in the network, the less dependent he is on relaying messages.

To find the shortest path from one node to all other nodes Floyd Warshall algorithm of all pairs shorted paths was applied to the adjacency matrix representing the nodes and links. This matrix gives a matrix of double, in which a numeric value represents a path between the node in the row to the node in the column and the value = measure of distance between them. The nodes which are disconnected have a value of ∞. Now closeness is calculated for every node using the sum of inverse methods.

Top 10 nodes having highest closeness

ID	34	74	33	16	30	65	58	76	26	68
Close	0.7772	0.7488	0.6558	0.6184	0.603	0.6027	0.6002	0.5906	0.5797	0.579



Farness measure as the opposite of closeness centrality

Top 10 people with highest farness

ID	116	122	4	1	5	0	21	105	93	121
Farness	615.0	546.0	493.0	478.0	472.0	470.0	444.0	423.0	421.0	418.0

Eigenvector Centrality

The distance and reach closeness measures are a function of all the distances from each node to the other nodes. Consider two nodes, one is central locally but remote with others and the other is at moderate distance from all other nodes. The eigenvector approach is used to find a central person in the global structure. The largest eigenvalues of the matrix are calculated, if they dominate the global structure, their corresponding eigenvectors will indicate the most important nodes. If the first eigenvalue occupies large percentage of the sum of all the eigenvalues and it's considerably larger than the second, its eigenvector should be checked.

Top five Eigen values in the Enron dataset

Eigen Values	20.65	11.12	9.50	9.08	8.52

Top 10 Nodes having highest Eigenvector centrality

ID	34	74	33	30	16	41	39	31	35

Betweenness Centrality

If two nodes are not adjacent but reachable, their geodesic must go through one or more nodes. The number of times a node lies on the geodesic of the nodes is the indicator of its centrality in the network. Thus, if a nodes lies on the geodesic of many nodes, its centrality would be higher.

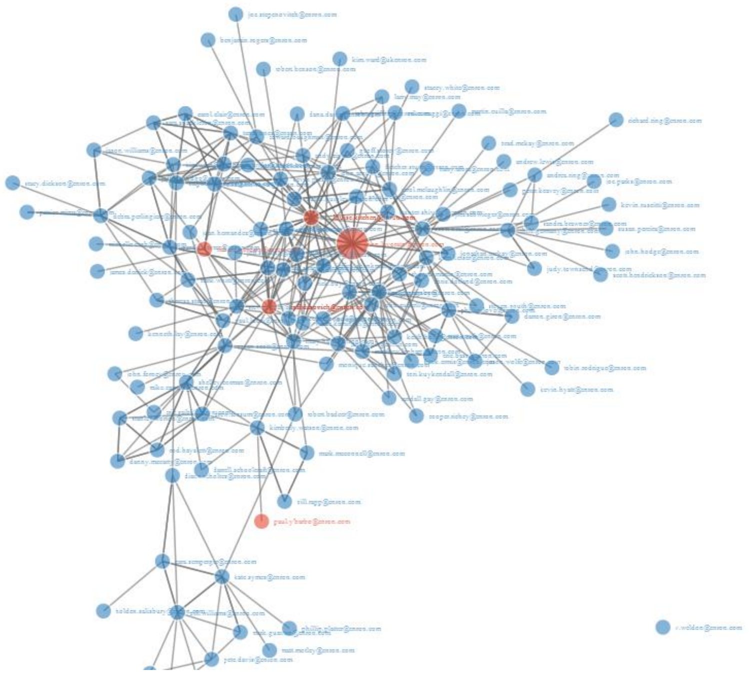
This centrality measure was calculated by implementing a faster algorithm for betweenness centrality given by Brandes called Brandes Algorithm for betweenness centrality.

References:http://www.markhneedham.com/blog/2013/07/19/graph-processing-calculating-betweenness-centrality-for-an-undirected-graph-using-graphstream/

http://algo.uni-konstanz.de/publications/b-fabc-01.pdf

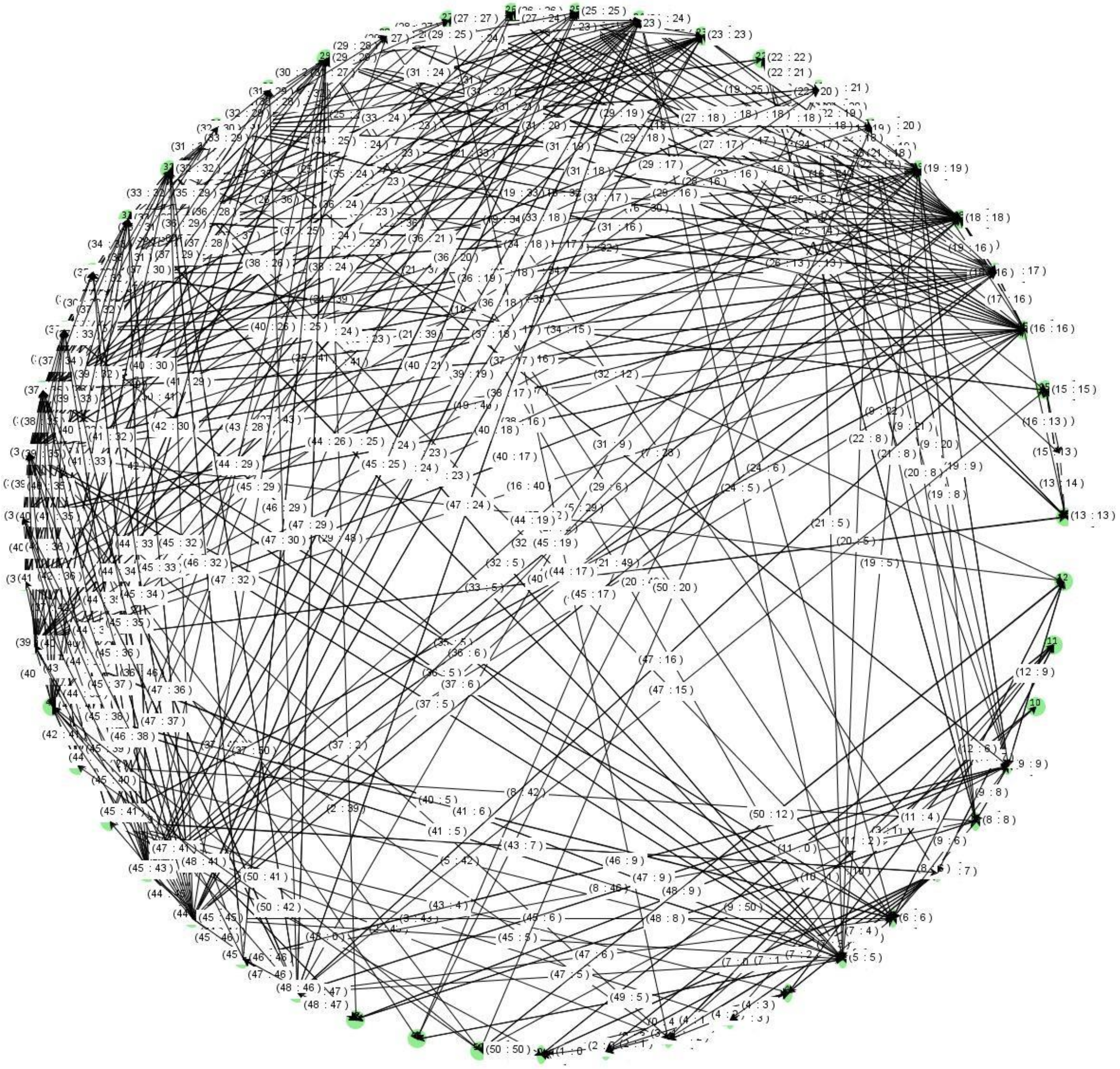
Top 10 nodes with most Betweenness

ID	34	92	58	74	33	2	16	65	71	76
Between	2855	1224	1050	1036	926	895	774	746	604	550

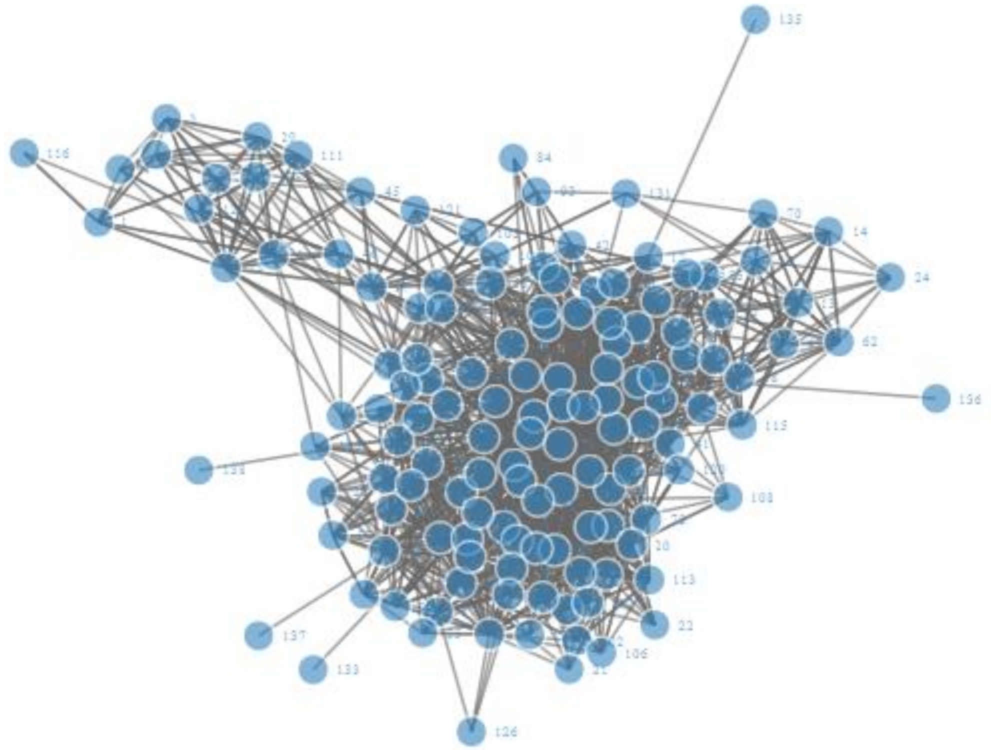


Transitivity

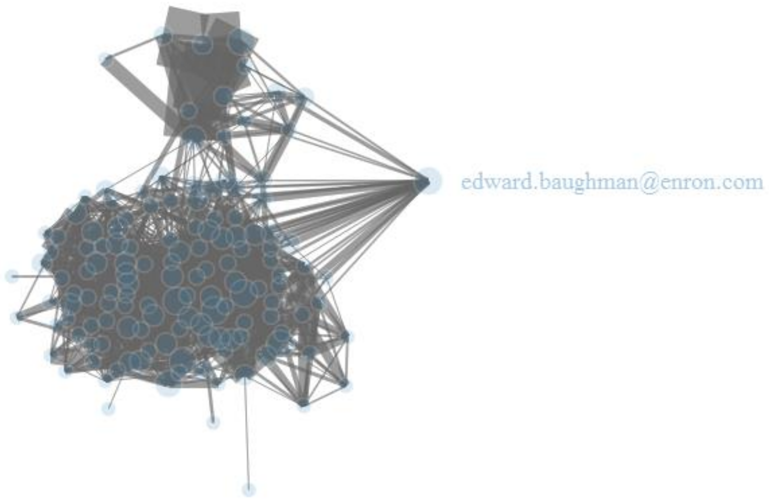
Transitivity is a property of the links in a network. If there is a tie from node u to v and v to y, then, in a transitive network node u and y would also be connected. It's calculated by determining the distances from a node to all others and returning a Boolean matrix having values "True" and "False". Value of "True" represents a connection from node in the row to the node in the column of matrix.



Enron Email data showing connections in between the top 50 nodes



150 nodes in the Enron network



Example of high Betweenness in the dataset

Mapping of Ids with Email Addresses of user

0	pete.davis@enron.com	10	mark.taylor@enron.com
1	ryan.slinger@enron.com	11	susan.bailey@enron.com
2	bill.williams@enron.com	12	carol.clair@enron.com
3	mark.guzman@enron.com	13	kimberly.watson@enron.com
4	craig.dean@enron.com	14	bill.rapp@enron.com
5	geir.solberg@enron.com	15	thomas.martin@enron.com
6	debra.perlingiere@enron.com	16	phillip.allen@enron.com
7	gerald.nemec@enron.com	17	joe.parks@enron.com
8	tana.jones@enron.com	18	rod.hayslett@enron.com
9	sara.shackleton@enron.com	60	jason.williams@enron.com
19	drew.fossum@enron.com	61	stephanie.panus@enron.com
20	chris.germany@enron.com	62	lynn.blair@enron.com
21	judy.townsend@enron.com	63	darron.giron@enron.com
22	scott.hendrickson@enron.com	64	stacy.dickson@enron.com
23	kate.symes@enron.com	65	kevin.presto@enron.com
24	paul.y'barbo@enron.com	66	dana.davis@enron.com
25	mark.mcconnell@enron.com	67	michelle.cash@enron.com
26	jeff.dasovich@enron.com	68	kam.keiser@enron.com
27	richard.shapiro@enron.com	69	phillip.love@enron.com
28	steven.kean@enron.com	70	darrell.schoolcraft@enron.com
29	holden.salisbury@enron.com	71	barry.tycholiz@enron.com
30	david.delainey@enron.com	72	fletcher.sturm@enron.com
31	jeff.skilling@enron.com	73	hunter.shively@enron.com
32	jeffrey.hodge@enron.com	74	sally.beck@enron.com
33	louise.kitchen@enron.com	75	martin.cuilla@enron.com
34	john.lavorato@enron.com	76	scott.neal@enron.com
35	greg.whalley@enron.com	77	larry.may@enron.com
36	matthew.lenhart@enron.com	78	sandra.brawner@enron.com
37	jay.reitmeyer@enron.com	79	brad.mckay@enron.com
38	john.zufferli@enron.com	80	jim.schwieger@enron.com
39	vince.kaminski@enron.com	81	peter.keavey@enron.com
40	stanley.horton@enron.com	82	dutch.quigley@enron.com
41	jeffrey.shankman@enron.com	83	andrea.ring@enron.com
42	steffes.james@enron.com	84	richard.ring@enron.com
43	richard.sanders@enron.com	85	shelley.corman@enron.com
44	robert.badeer@enron.com	86	rick.buy@enron.com
45	chris.stokley@enron.com	87	eric.bass@enron.com
46	diana.scholtes@enron.com	88	jane.tholt@enron.com

47	lindy.donoho@enron.com	89	randall.gay@enron.com
48	mike.mcconnell@enron.com	90	patrice.mims@enron.com
49	jason.wolfe@enron.com	91	paul.lucci@enron.com
50	john.arnold@enron.com	92	edward.baughman@enron.com
51	mike.maggi@enron.com	93	joe.stepenovitch@enron.com
52	andy.zipper@enron.com	94	tori.kuykendall@enron.com
53	elizabeth.sager@enron.com	95	james.derrick@enron.com
54	keith.holst@enron.com	96	mark.haedicke@enron.com
55	robin.rodrigue@enron.com	97	john.forney@enron.com
56	errol.mclaughlin@enron.com	98	mike.carson@enron.com
57	frank.ermis@enron.com	99	matt.smith@enron.com
58	susan.scott@enron.com	134	mary.fischer@enron.com
59	kim.ward@enron.com	135	kay.mann@enron.com
100	theresa.staab@enron.com	136	kenneth.lay@enron.com
101	mark.whitt@enron.com	137	judy.hernandez@enron.com
102	vladi.pimenov@enron.com	138	mike.swerzbin2@enron.com
103	geoff.storey@enron.com		
104	eric.saibi@enron.com		
105	juan.hernandez@enron.com		
106	john.hodge@enron.com		
107	jonathan.mckay@enron.com		
108	susan.pereira@enron.com		
109	benjamin.rogers@enron.com		
110	harry.arora@enron.com		
111	cara.semperger@enron.com		
112	kevin.ruscitti@enron.com		
113	john.griffith@enron.com		
114	monique.sanchez@enron.com		
115	danny.mccarty@enron.com		
116	steven.merris@enron.com		
117	matt.motley@enron.com		
118	robert.benson@enron.com		
119	chris.dorland@enron.com		
120	tom.donohoe@enron.com		
121	clint.dean@enron.com		
122	steven.south@enron.com		
123	andrew.lewis@enron.com		
124	joe.quenet@enron.com		
125	phillip.platter@enron.com		
126	kevin.hyatt@enron.com		

127	lisa.gang@enron.com
128	cooper.richey@enron.com
129	v.weldon@enron.com
130	stacey.white@enron.com
131	larry.campbell@enron.com
132	kim.ward@ukenron.com
133	m..scott@enron.com