Graph Mining

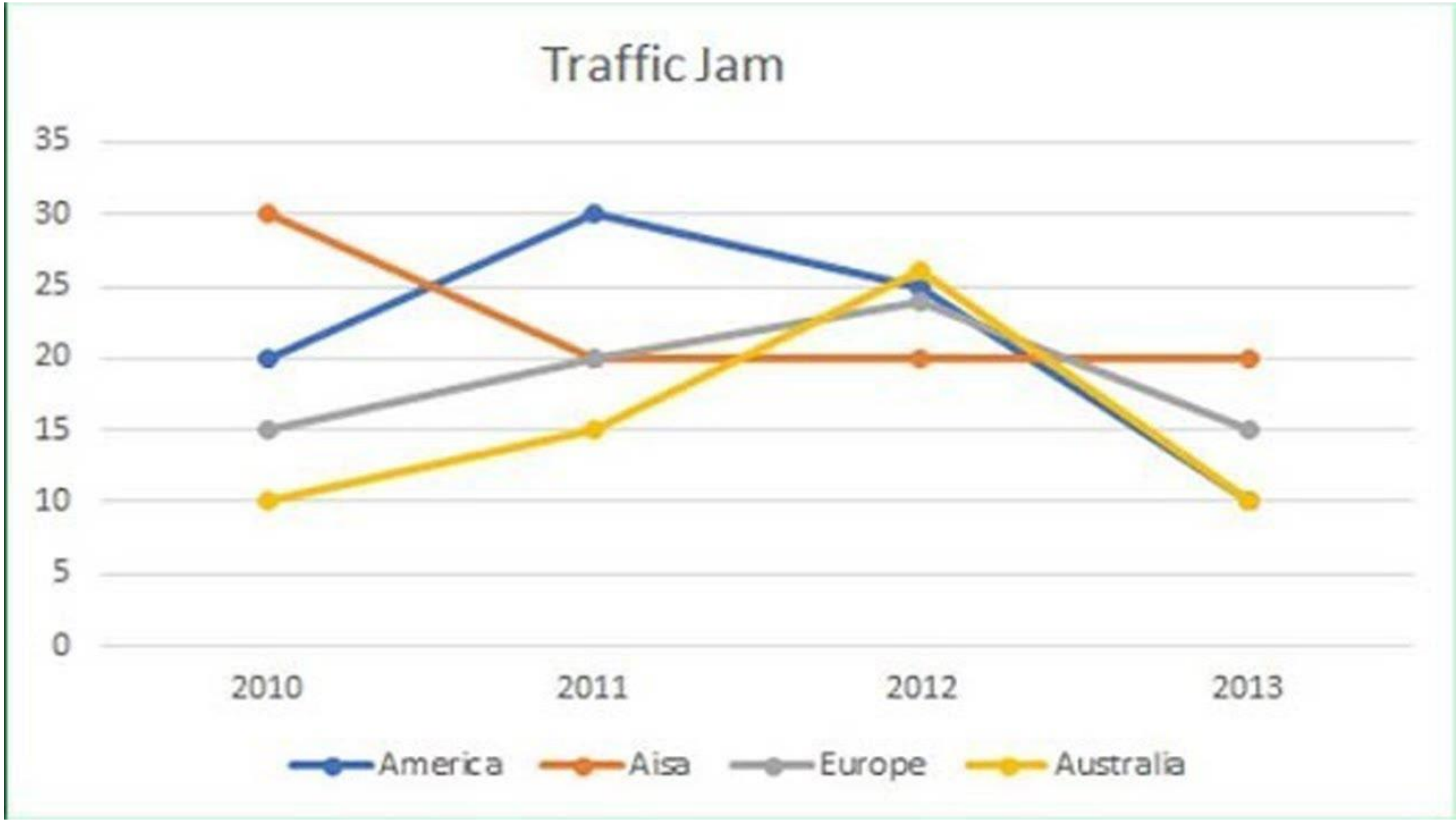
Dr. Hardik Joshi

Department of Computer Sc.

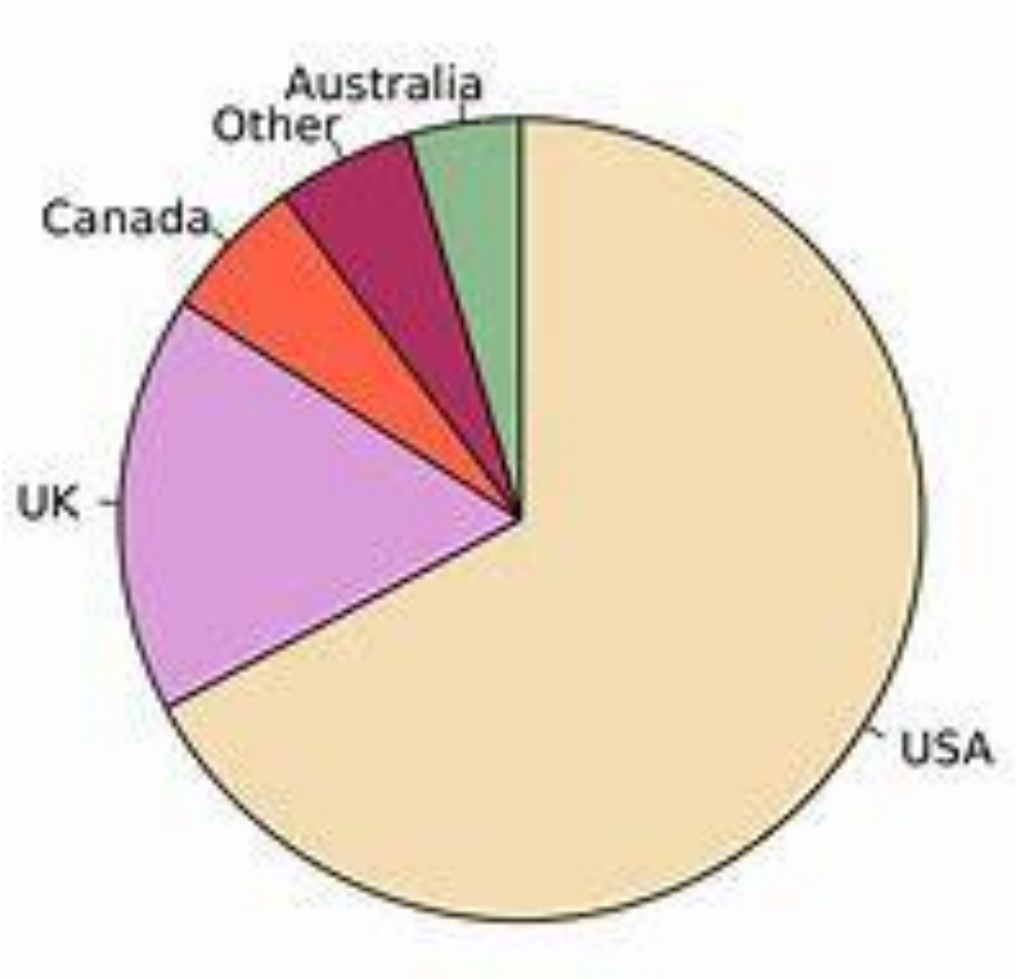
Gujarat University

E-mail: hardikjoshi@gujaratuniversity.ac.in

Is this a Graph?

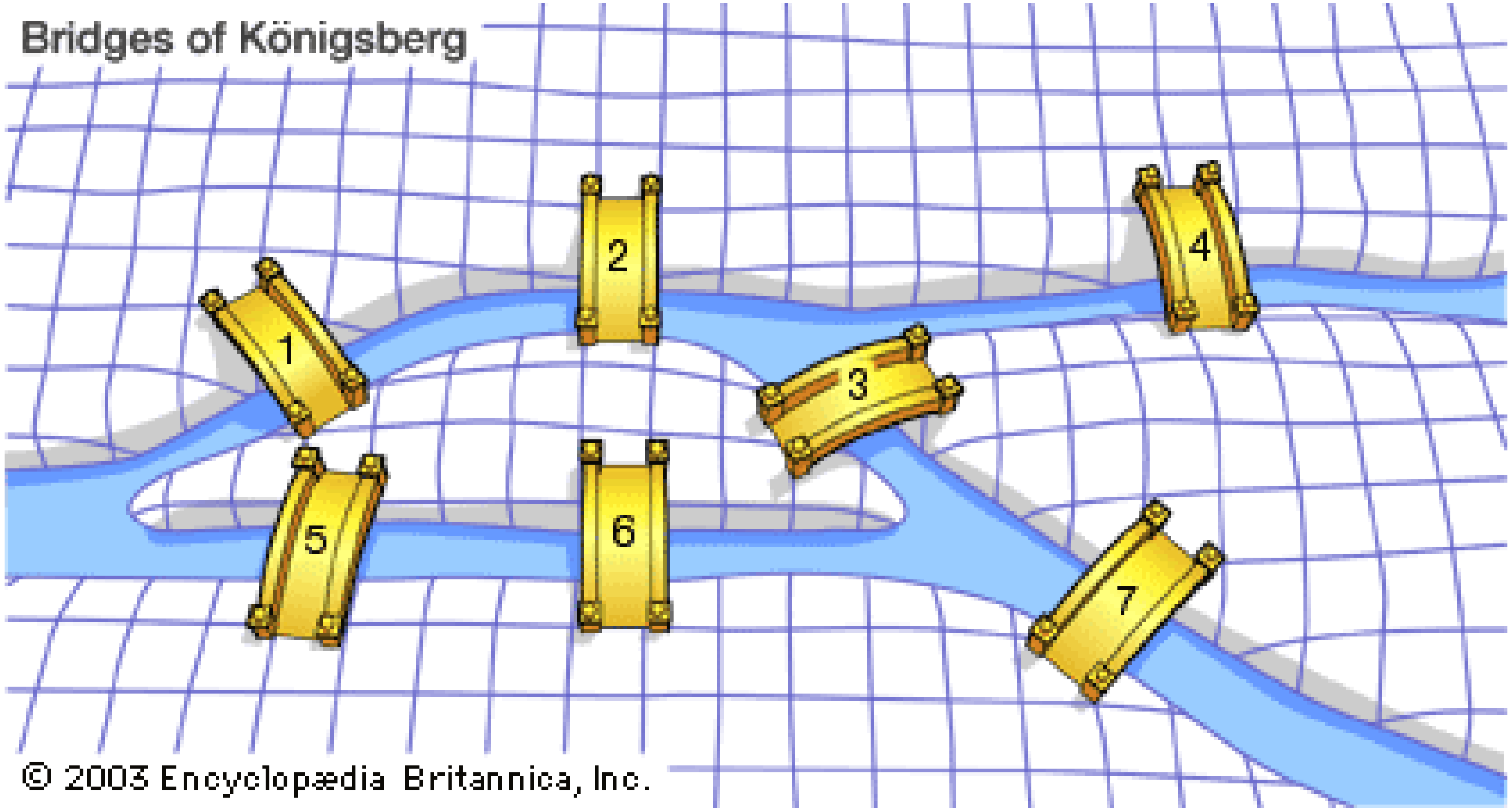


Is this a Graph?

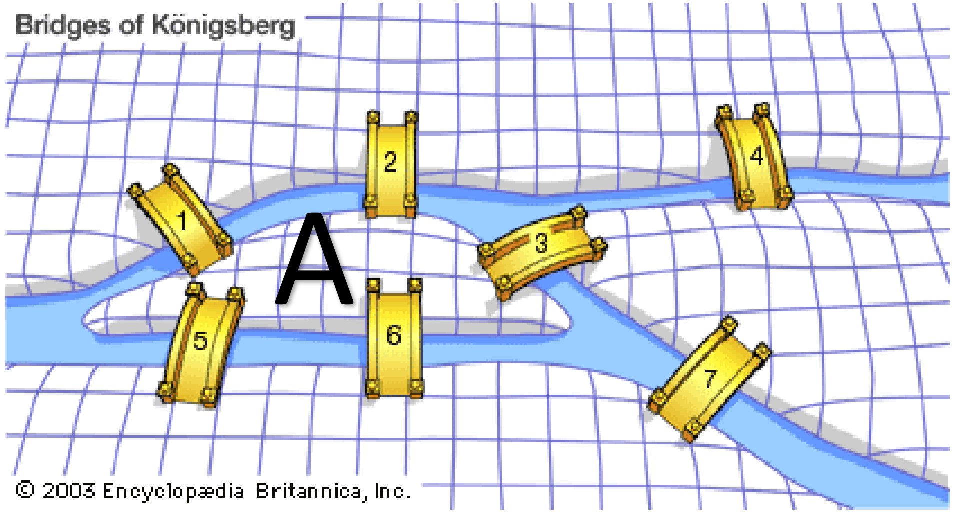


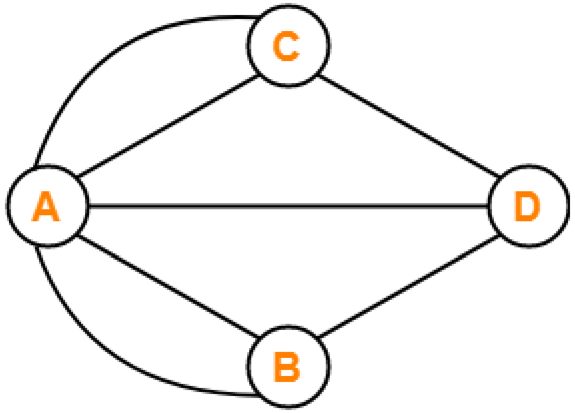
History of Graph Theory

History of Graph Theory



History of Graph Theory

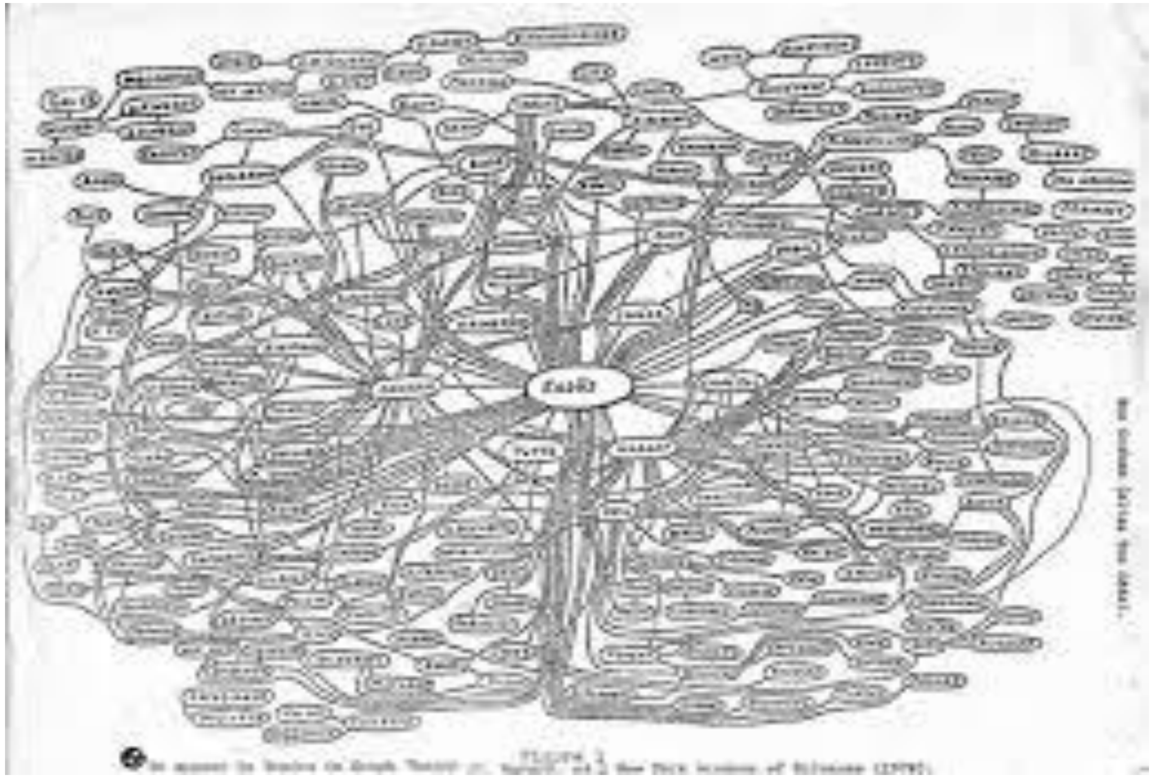




Graph Representation

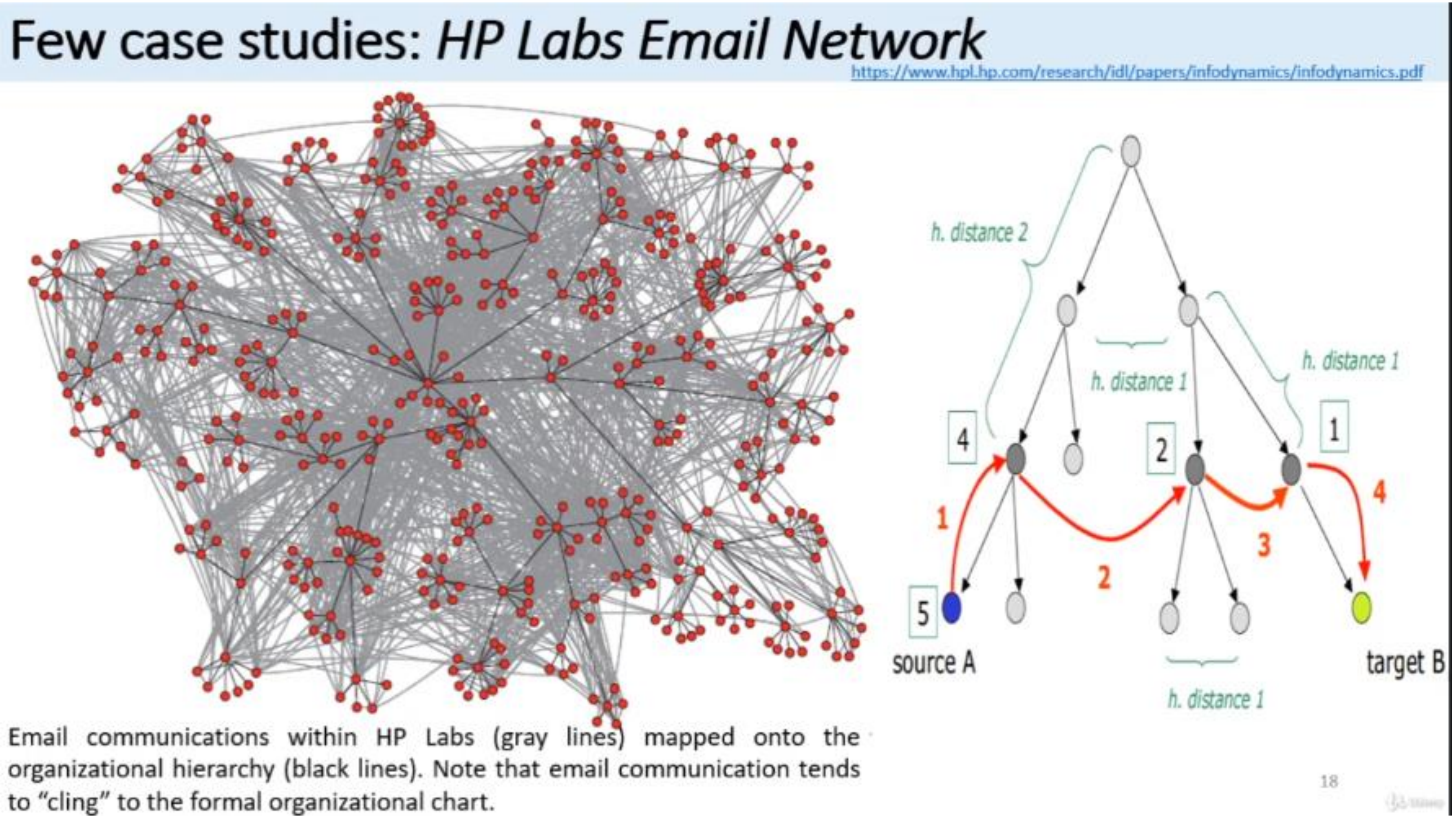
History of Graph Theory
(Euler & Dijkstra)

Erdos Number



The **Erdős number** (Hungarian: [ˈɛrdøːʃ]) describes the "collaborative distance" between mathematician Paul **Erdős** and another person, as measured by authorship of mathematical papers.

Why Graph Analytics can be helpful ?

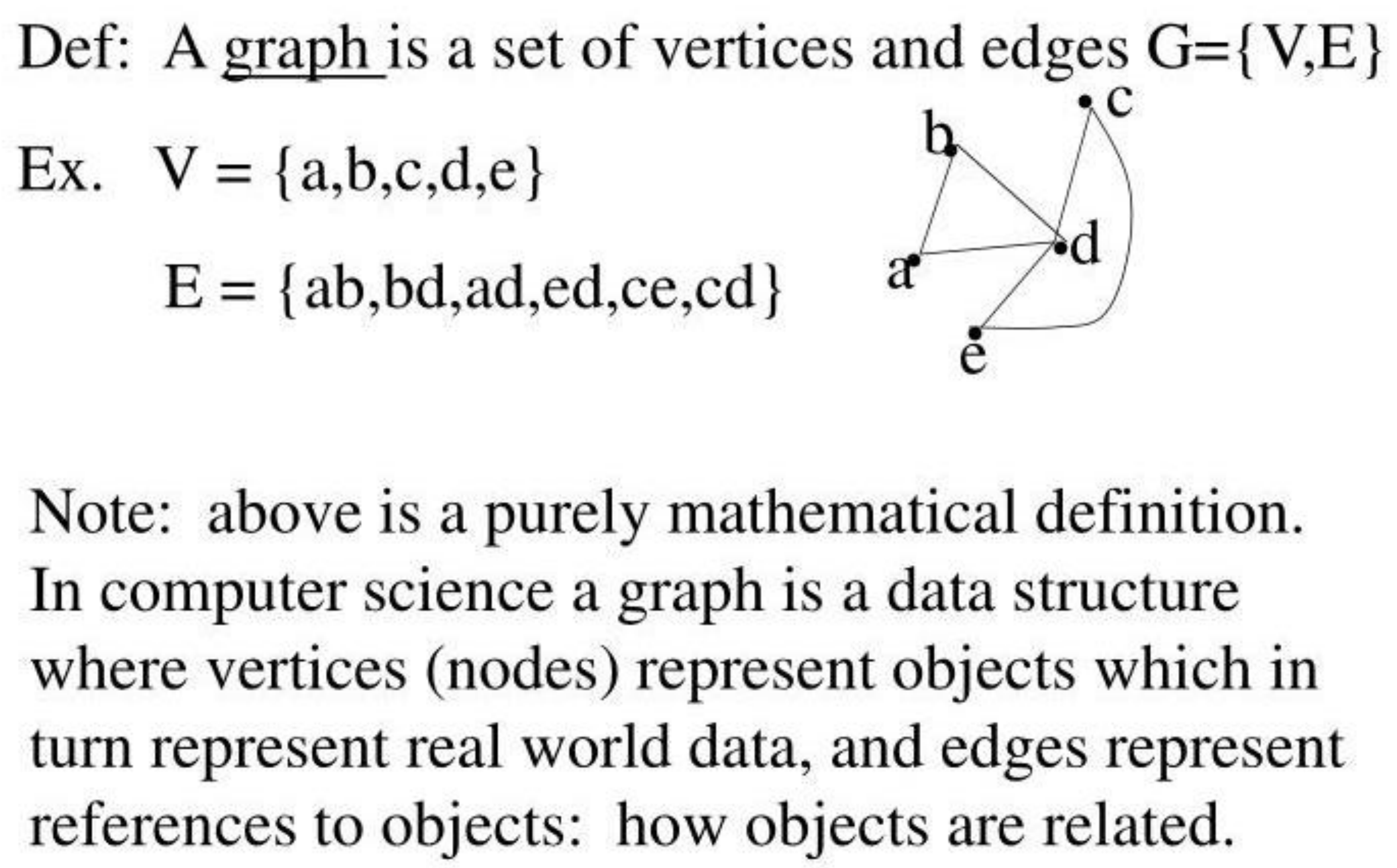


Few case studies: *HP Labs Email Network*

https://www.hpl.hp.com/research/idl/papers/infodynamics/infodynamics.pdf

Email communications within HP Labs (gray lines) mapped onto the organizational hierarchy (black lines). Note that email communication tends to "cling" to the formal organizational chart.

Defining a Graph



Def: A graph is a set of vertices and edges G={V,E}

Ex. V = {a,b,c,d,e}

E = {ab,bd,ad,ed,ce,cd}

Note: above is a purely mathematical definition. In computer science a graph is a data structure where vertices (nodes) represent objects which in turn represent real world data, and edges represent references to objects: how objects are related.

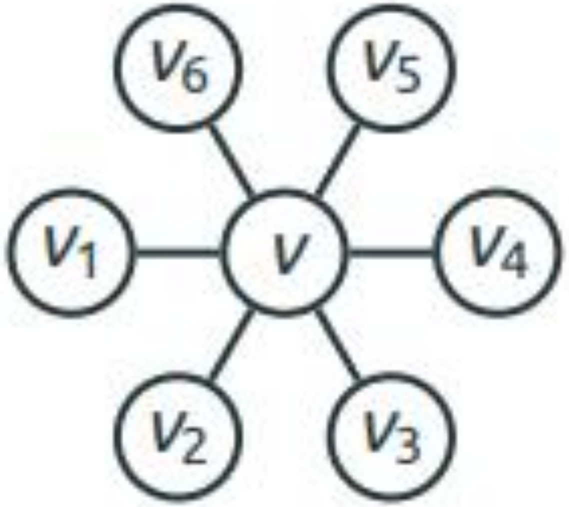
The Degree of a Vertex

- The Degree of a vertex is the number of its incident edges
- I.e., the Degree of a vertex is the number of its neighbors
- The degree of a vertex v is denoted by deg(v)
- The degree of a graph is the maximum degree of its vertices

The Degree of a Vertex: Examples

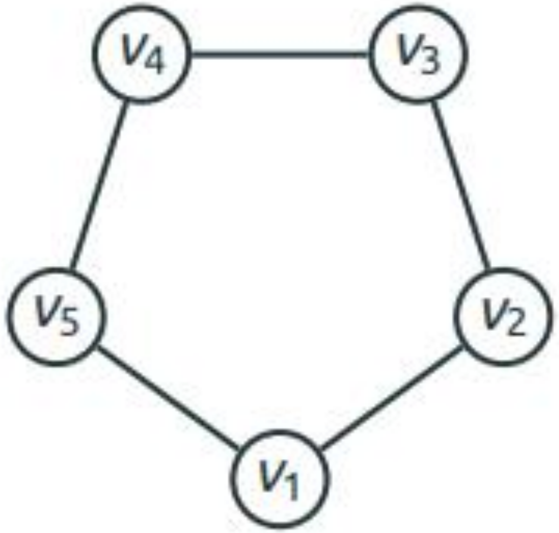
The degree of v is 6: $\deg(v) = 6$

The degree of v_6 is 1: $\deg(v_6) = 1$

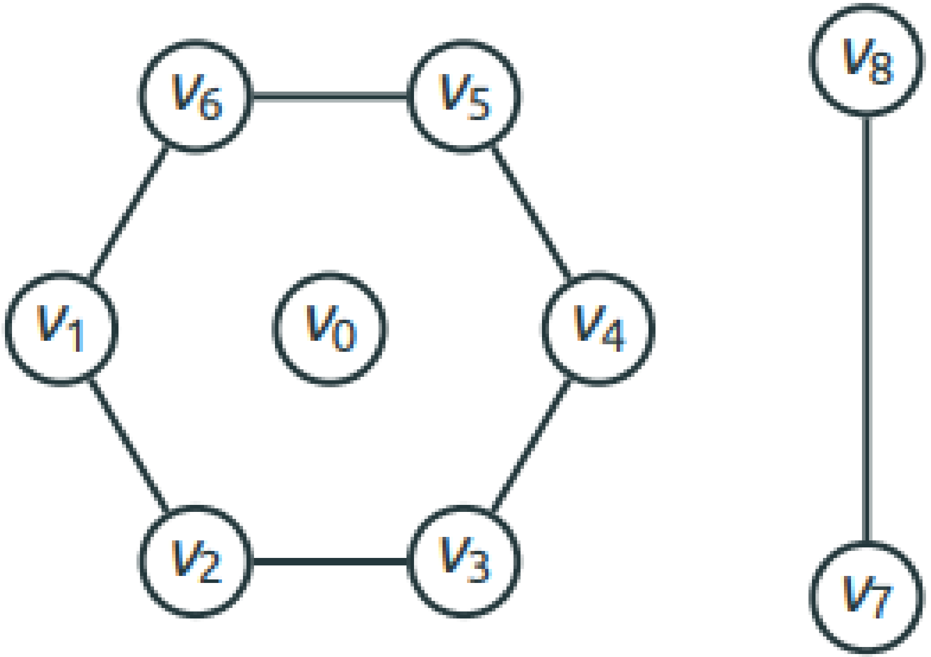


The Degree of a Vertex: Examples

The degree of every vertex is 2: $\forall i, \deg(v_i) = 2$

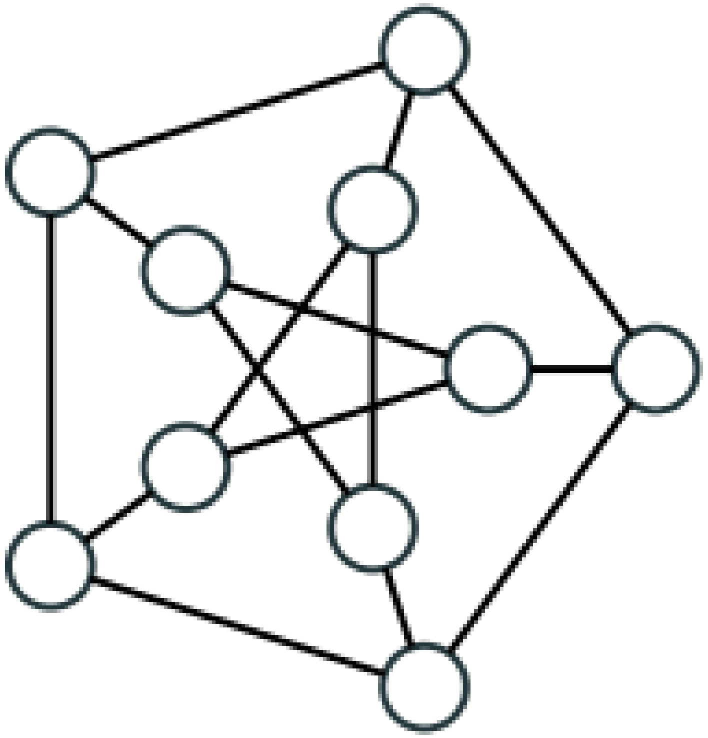


Isolated Vertices



Regular Graphs

A Regular graph is a graph where each vertex has the same degree

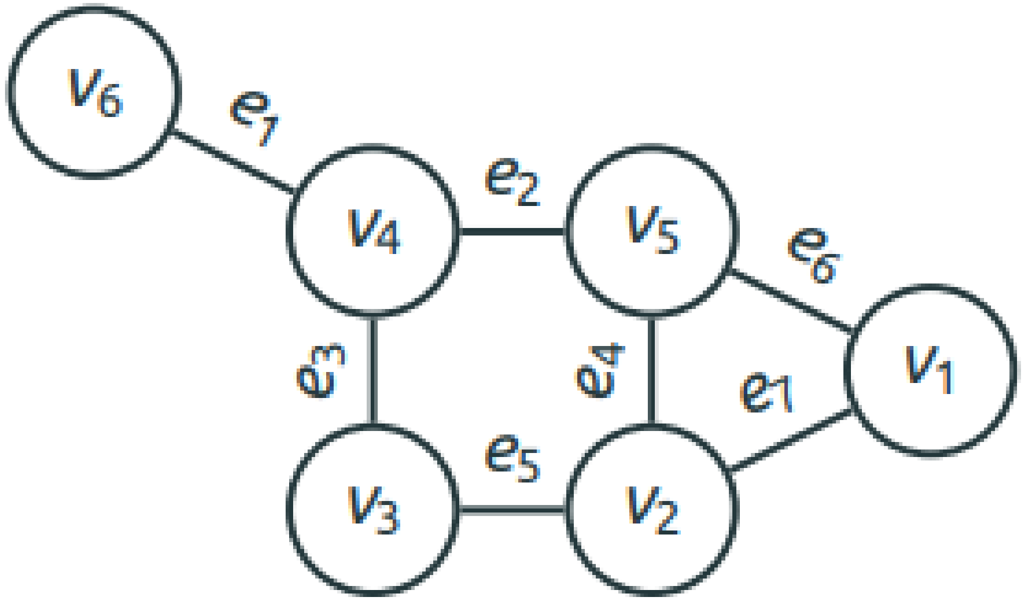


Walks

- A Walk in a graph is a sequence of edges, such that each edge (except for the first one) starts with a vertex where the previous edge ended
- The Length of a walk is the number of edges in it
- A Path is a walk where all edges are distinct
- A Simple Path is a walk where all vertices are distinct

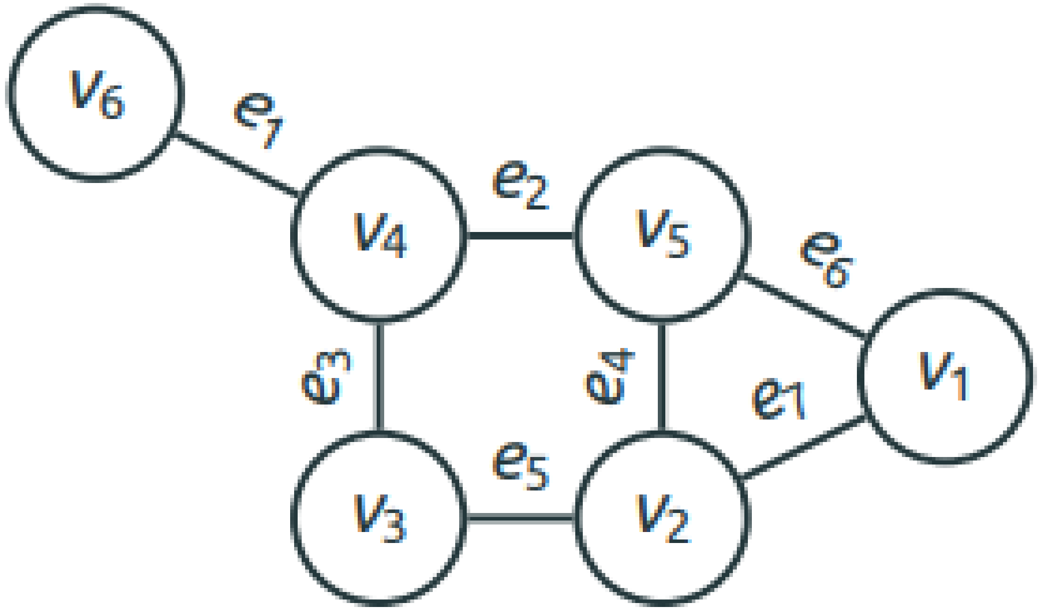
Walks: Examples

A walk of length 6: $(e_1, e_2, e_4, e_5, e_3, e_1)$
Not a path: uses e_1 twice



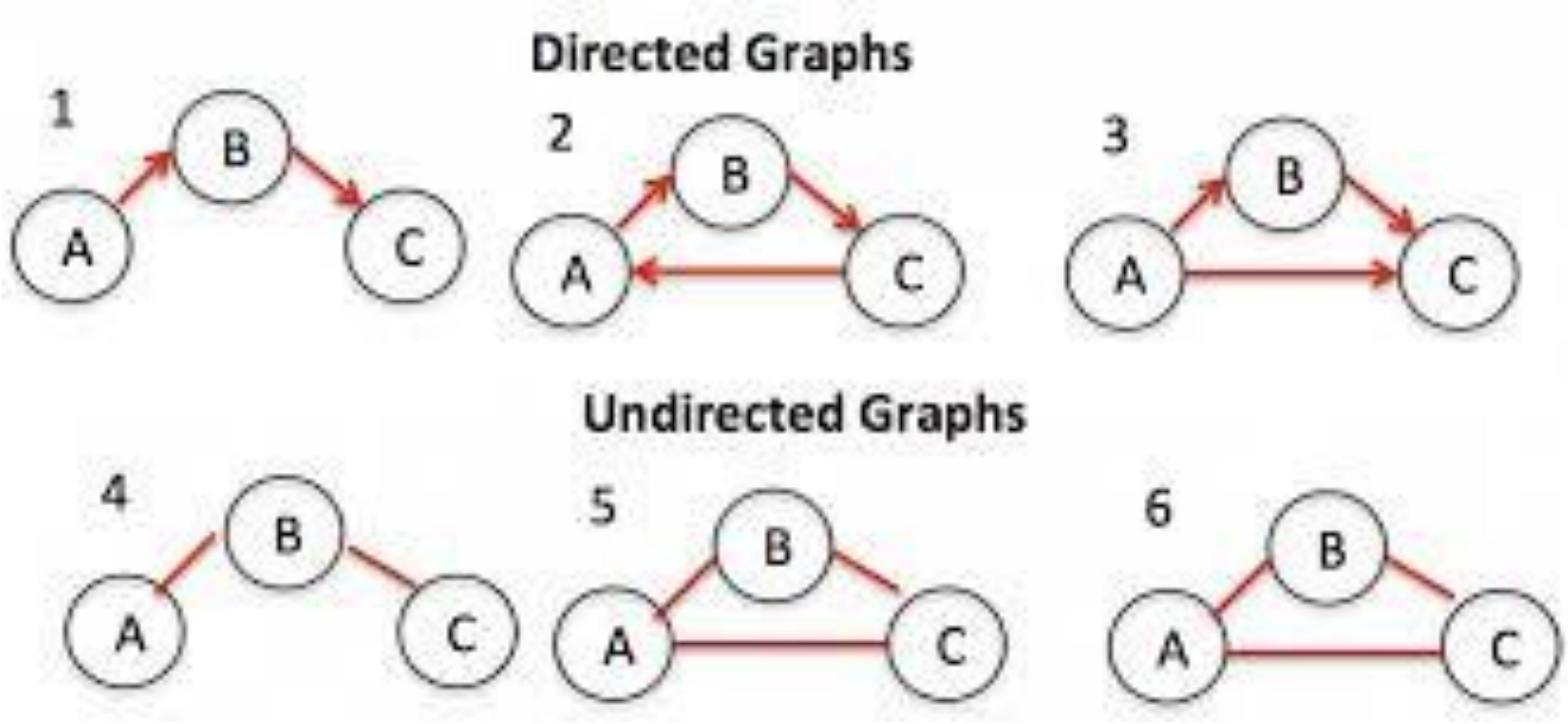
Walks: Examples

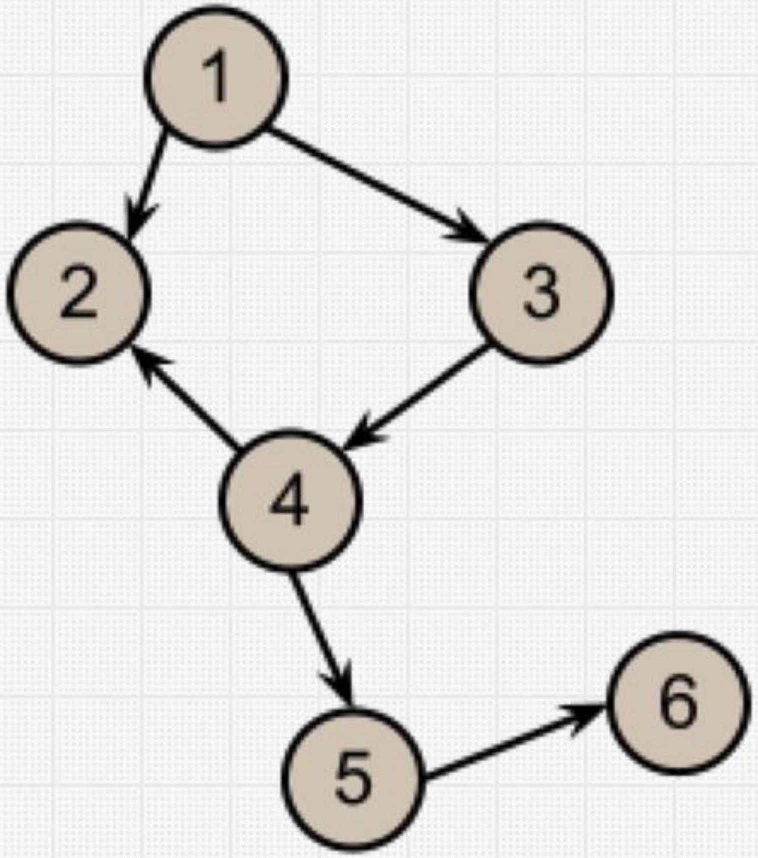
A path of length 4: (e_7, e_6, e_4, e_5)



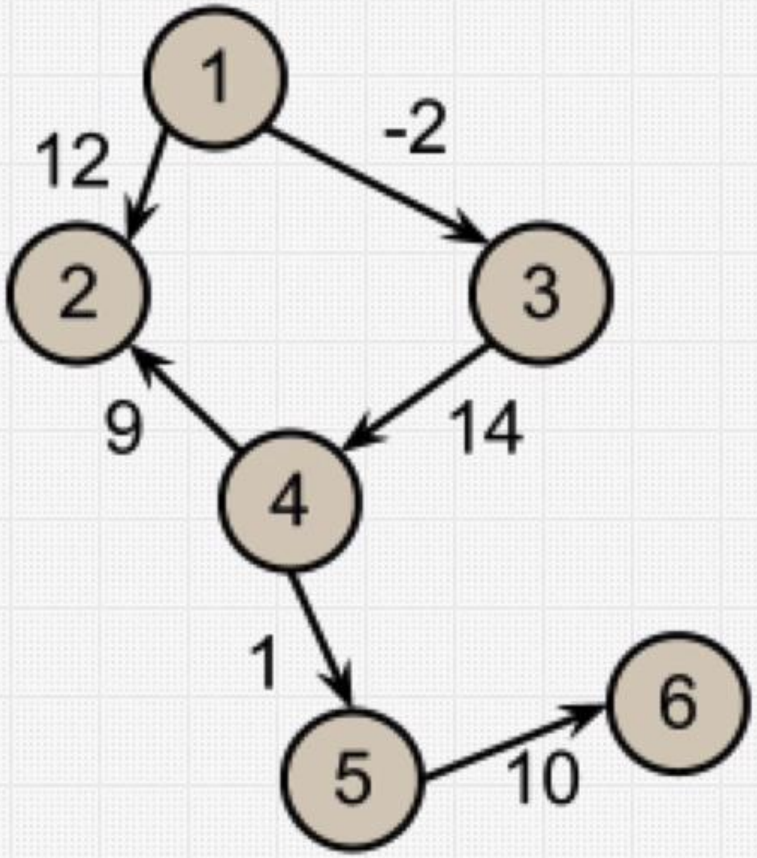
Cycles

- A Cycle in a graph is a path whose first vertex is the same as the last one
- In particular, all the edges in a Cycle are distinct
- A Simple Cycle is a cycle where all vertices except for the first one are distinct. (And there first vertex is taken twice)


Directed Graphs
1
A
B
C
2
A
B
C
3
A
B
C
Undirected Graphs
4
A
B
C
5
A
B
C
6
A
B
C

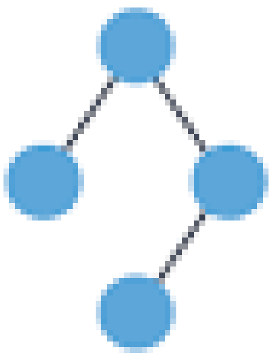


Unweighted

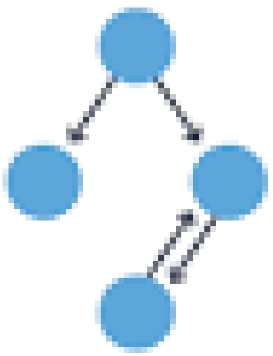


Weighted

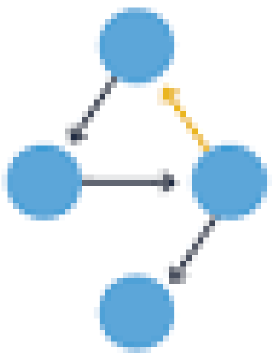
Undirected



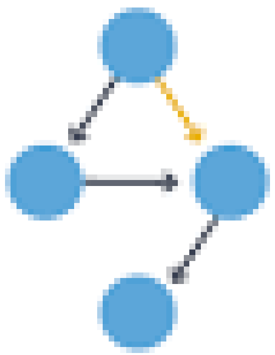
Directed



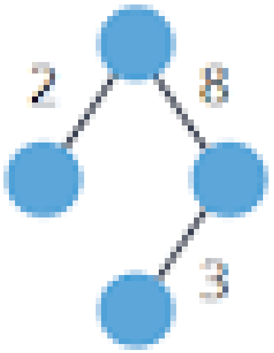
Cyclic



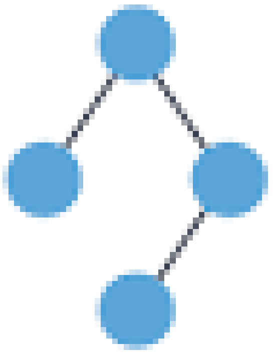
Acyclic



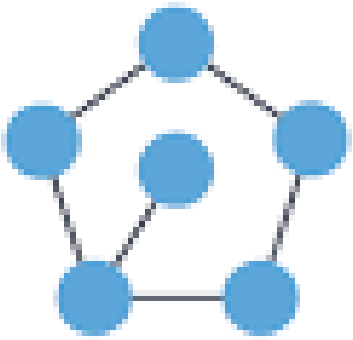
Weighted



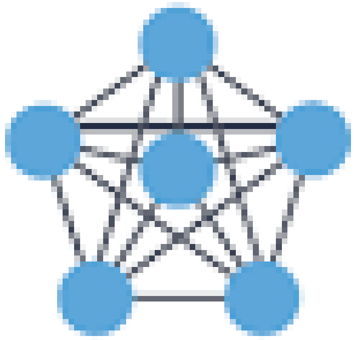
Unweighted



Sparse



Dense



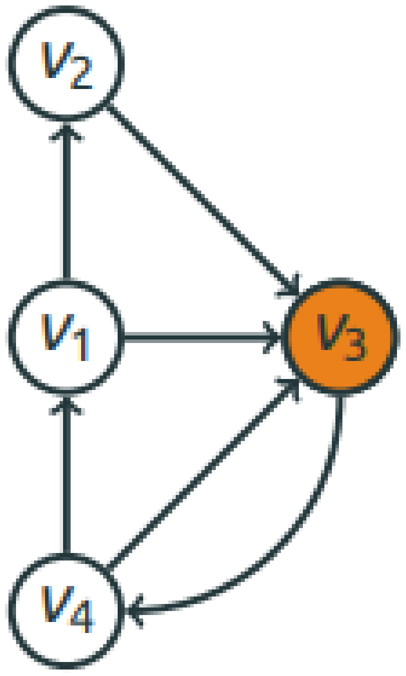
The Degree of a Vertex

- The Indegree of a vertex *v* is the number of edges ending at *v*

- The Outdegree of a vertex *v* is the number of edges leaving *v*

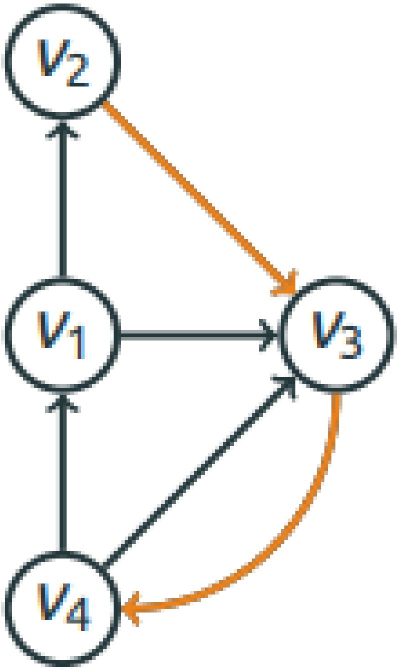
The Degree of a Vertex: Examples

The Indegree of v_3 is 3,
the Outdegree of v_3 is 1



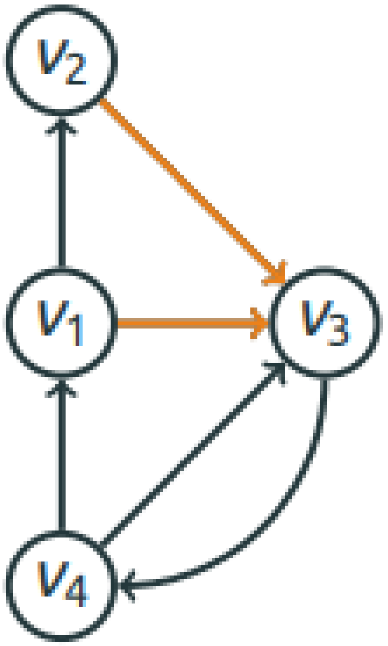
Directed Paths

(v_2, v_3, v_4) is a Path of length 2



Directed Paths

(v_1, v_3, v_2) is not a Path

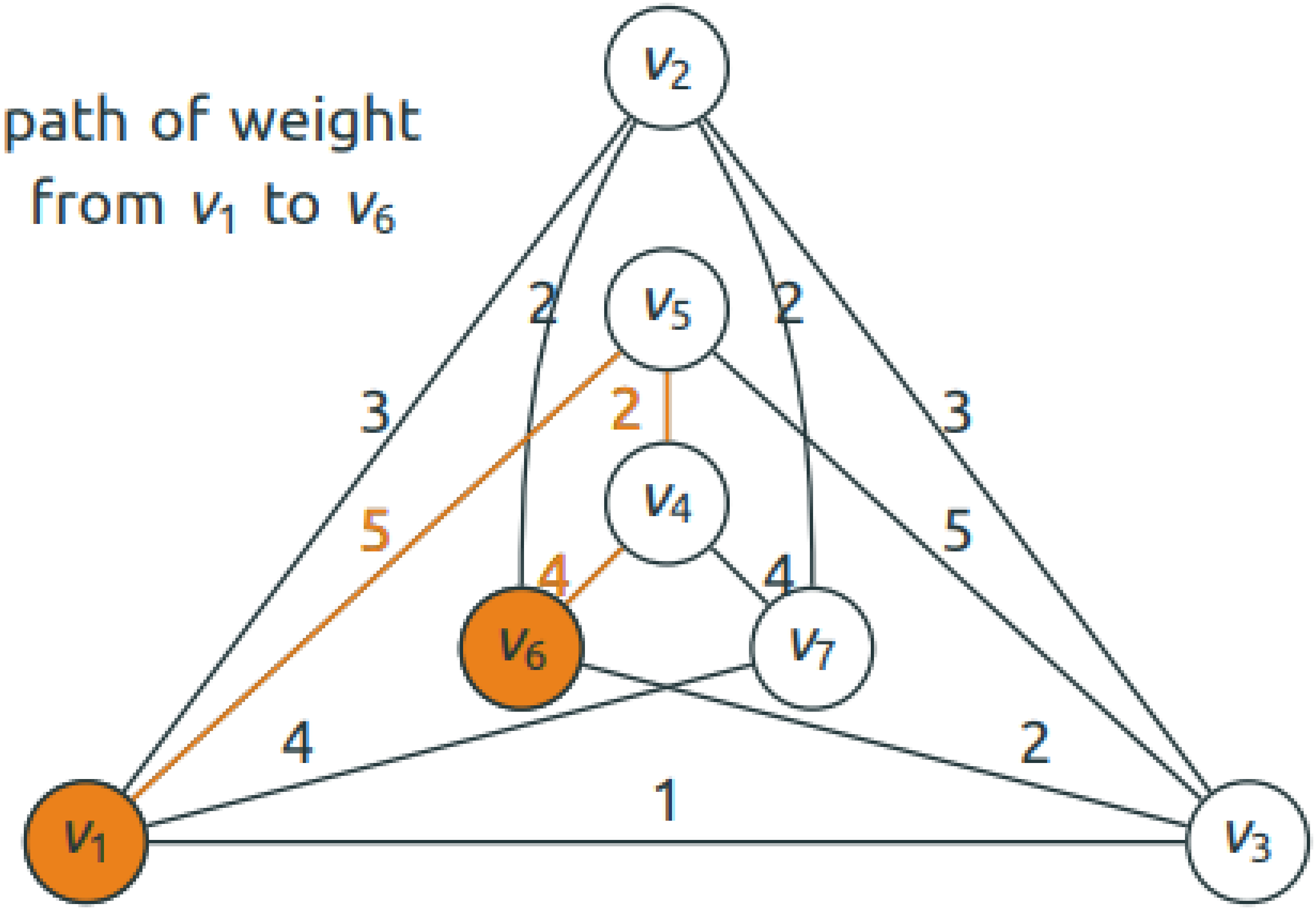


Weighted Paths

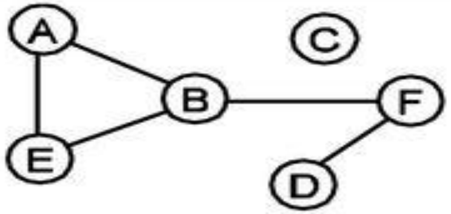
- A Weighted Graph associates a weight with every edge
- The Weight of a path is the sum of the weights of its edges
- A Shortest Path between two vertices is a path of the minimum weight
- The Distance between two vertices is the length of a shortest path between them

Weighted Paths: Examples

A path of weight 11 from v_1 to v_6

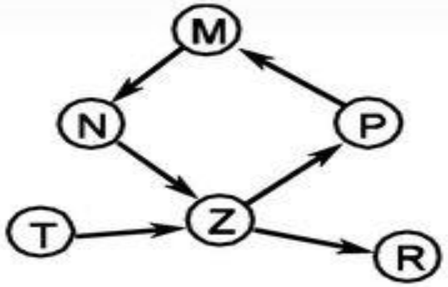


Adjacency Matrix Representation



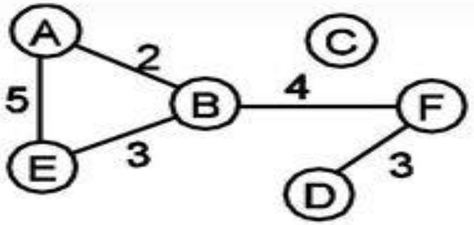
	A	B	C	D	E	F
A	0	1	0	0	1	0
B	1	0	0	0	1	1
C	0	0	0	0	0	0
D	0	0	0	0	0	1
E	1	1	0	0	0	0
F	0	1	0	1	0	0

(a) Adjacency matrix for an undirected graph.



	M	N	P	R	T	Z
M	0	1	0	0	0	0
N	0	0	0	0	0	1
P	1	0	0	0	0	0
R	0	0	0	0	0	0
T	0	0	0	0	0	1
Z	0	0	1	1	0	0

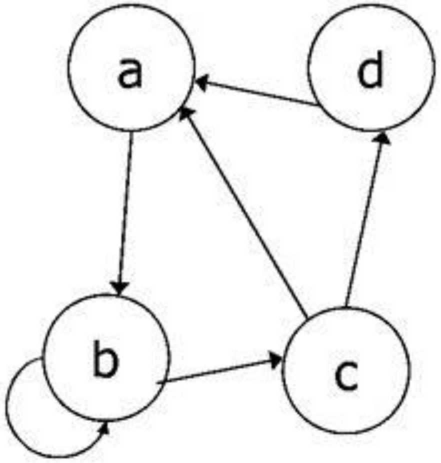
(b) Adjacency matrix for a directed graph.



	A	B	C	D	E	F
A	0	2	0	0	5	0
B	2	0	0	0	3	4
C	0	0	0	0	0	0
D	0	0	0	0	0	3
E	5	3	0	0	0	0
F	0	4	0	3	0	0

(c) Adjacency matrix for an undirected weighted graph.

Representing Graphs



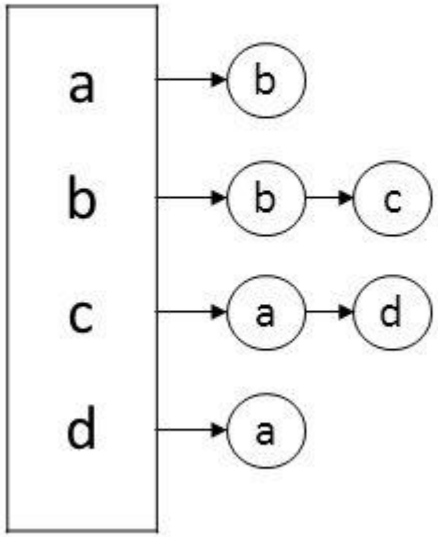
- Directed, unweighted

Adjacency matrix

source	a	b	c	d
a	0	1	0	0
b	0	1	1	0
c	1	0	0	1
d	1	0	0	0

target

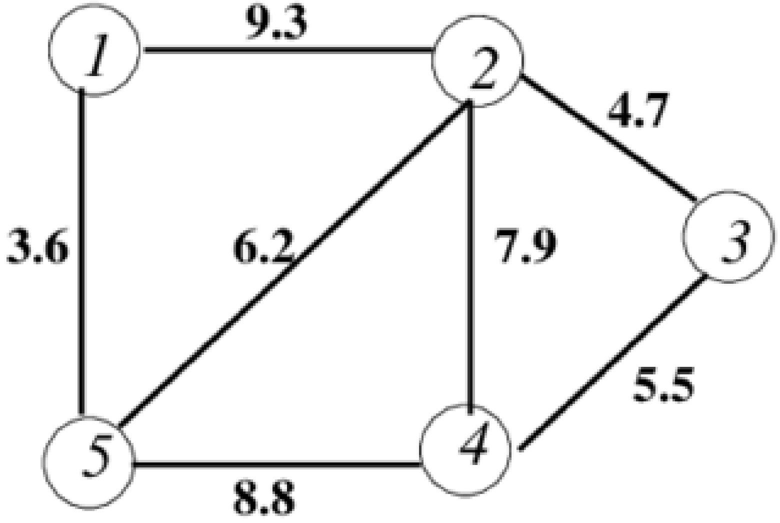
Adjacency List


a → b

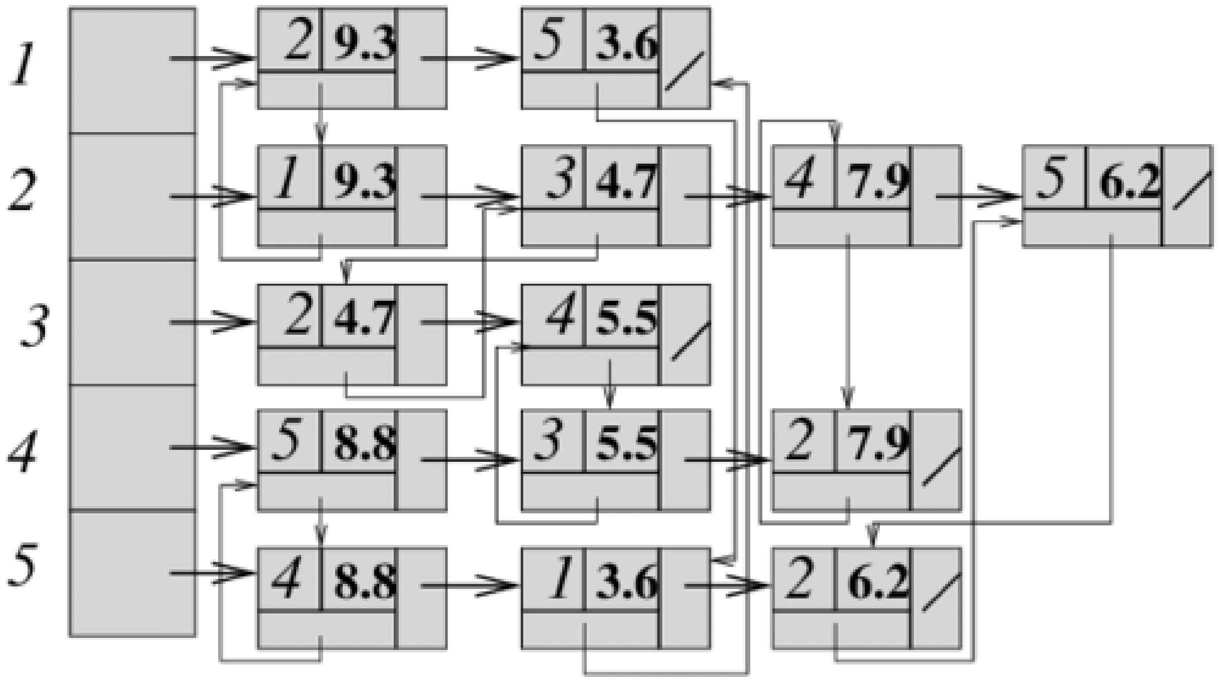
b → b → c

c → a → d

d → a


1
2
3
4
5
9.3
4.7
3.6
6.2
7.9
5.5
8.8

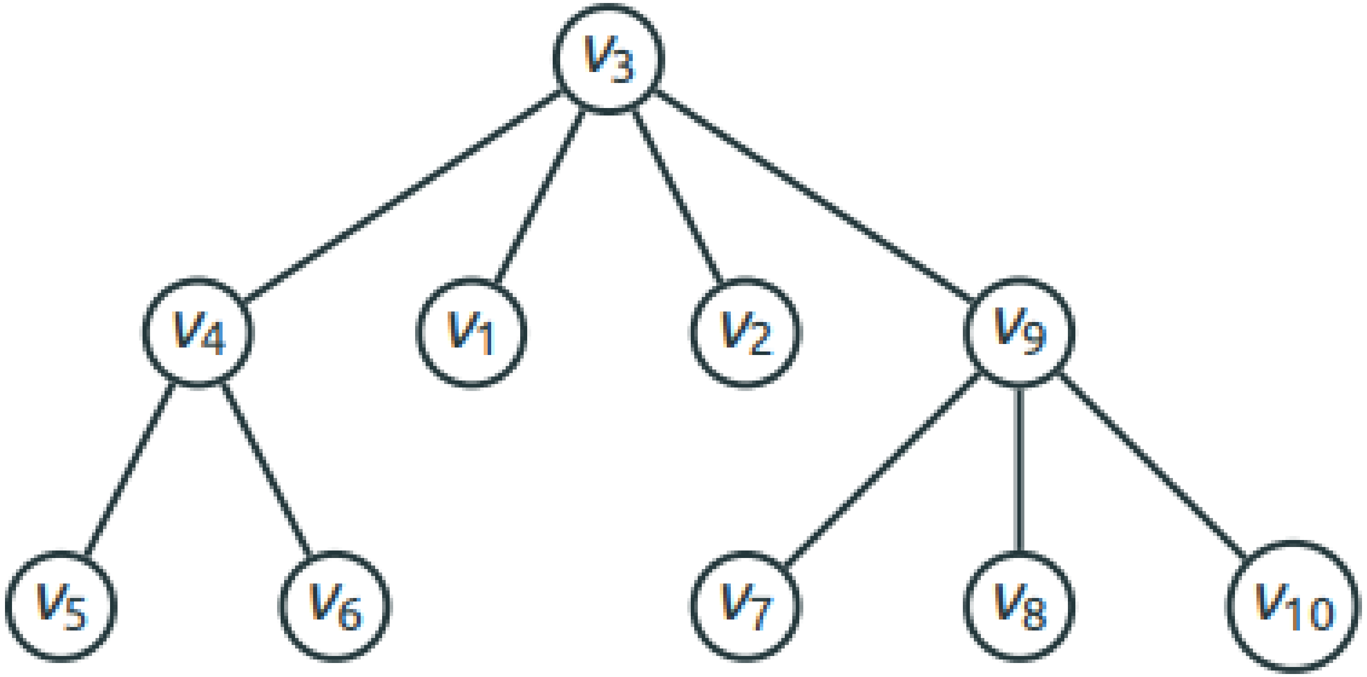
(a)


1
2
3
4
5
2 9.3
5 3.6
1 9.3
3 4.7
4 7.9
5 6.2
2 4.7
4 5.5
5 8.8
3 5.5
2 7.9
4 8.8
1 3.6
2 6.2

(b)

Drawing a Tree

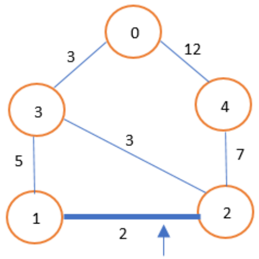
Connected; the number of edges is $n - 1$



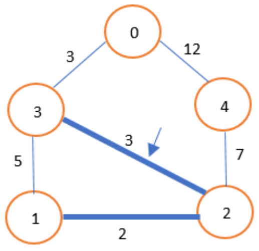
Definition

- A tree is a connected graph without cycles
- A tree is a connected graph on n vertices with $n - 1$ edges
- A graph is a tree if and only if there is a unique simple path between any pair of its vertices

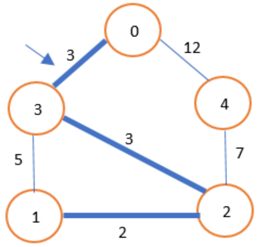
Processing Graphs (Data Structures)



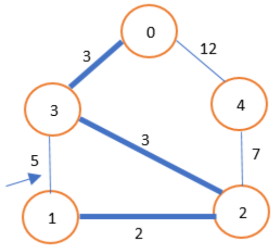
1



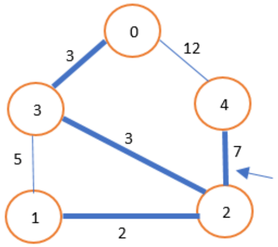
2



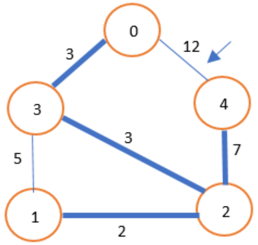
3



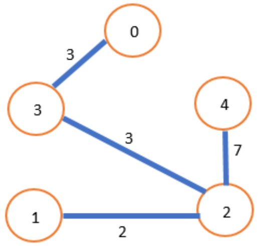
4



5



6

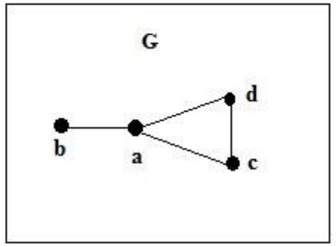


7

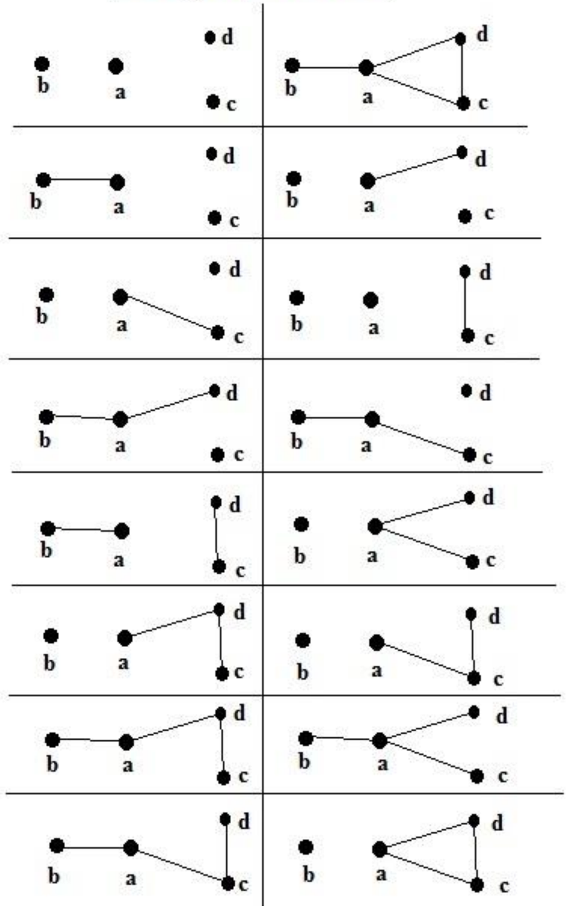
Algorithms:

- Prims Algorithm for MST
- Kruskal’s Algorithm for MST
- Shortest Path Algorithm
- Depth First Search
- Breadth First Search

Processing Graphs & Subgraphs

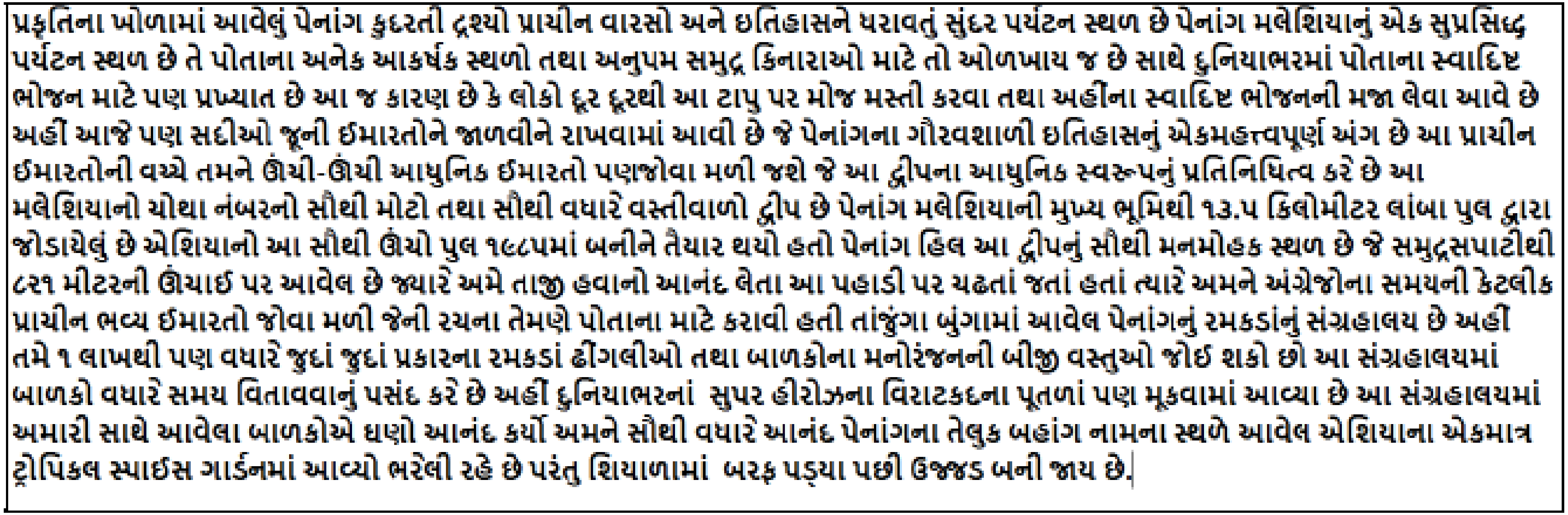


Subgraphs of G



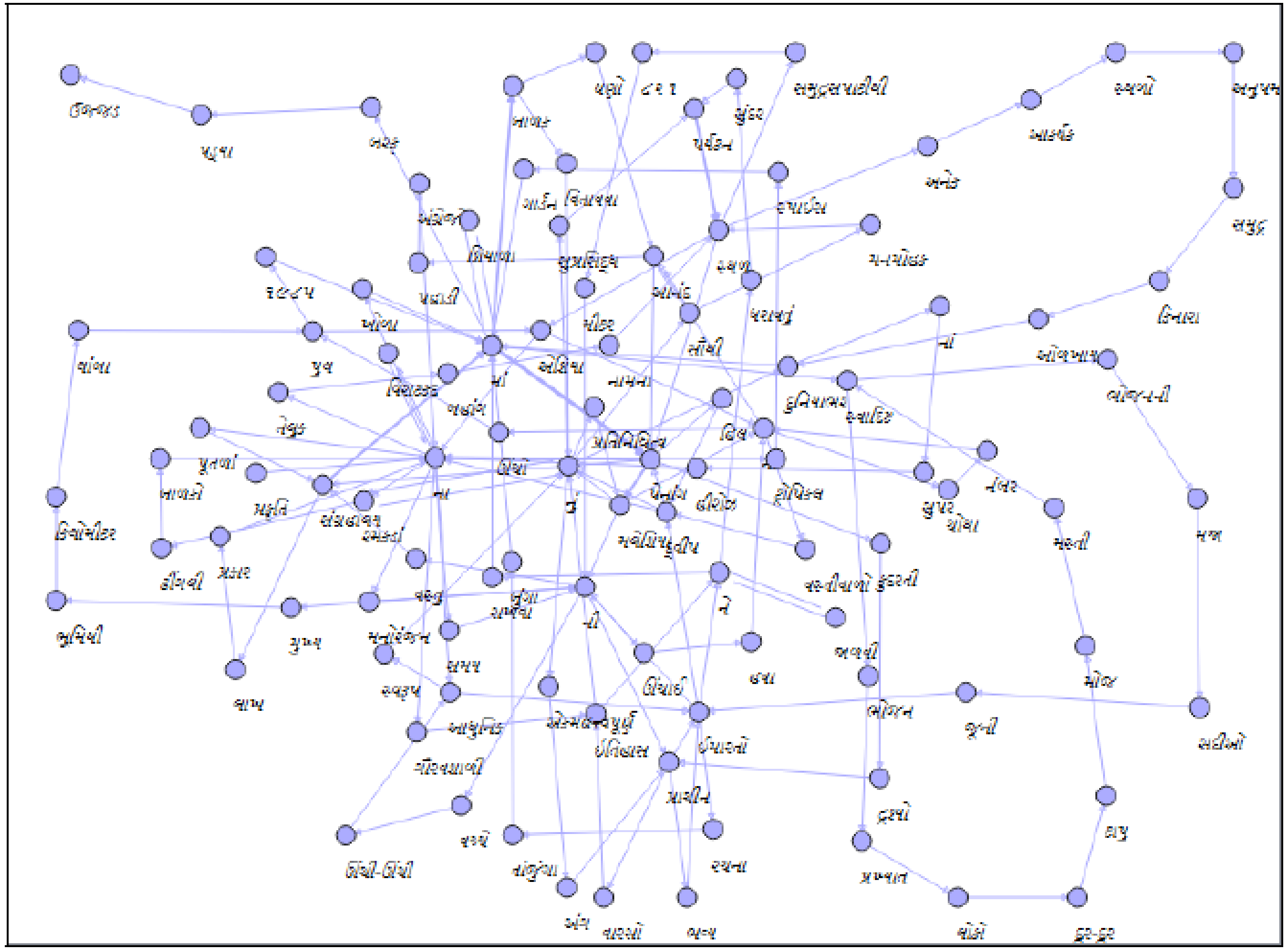
Case - Study

Text Graphs (Word co-occurrence)

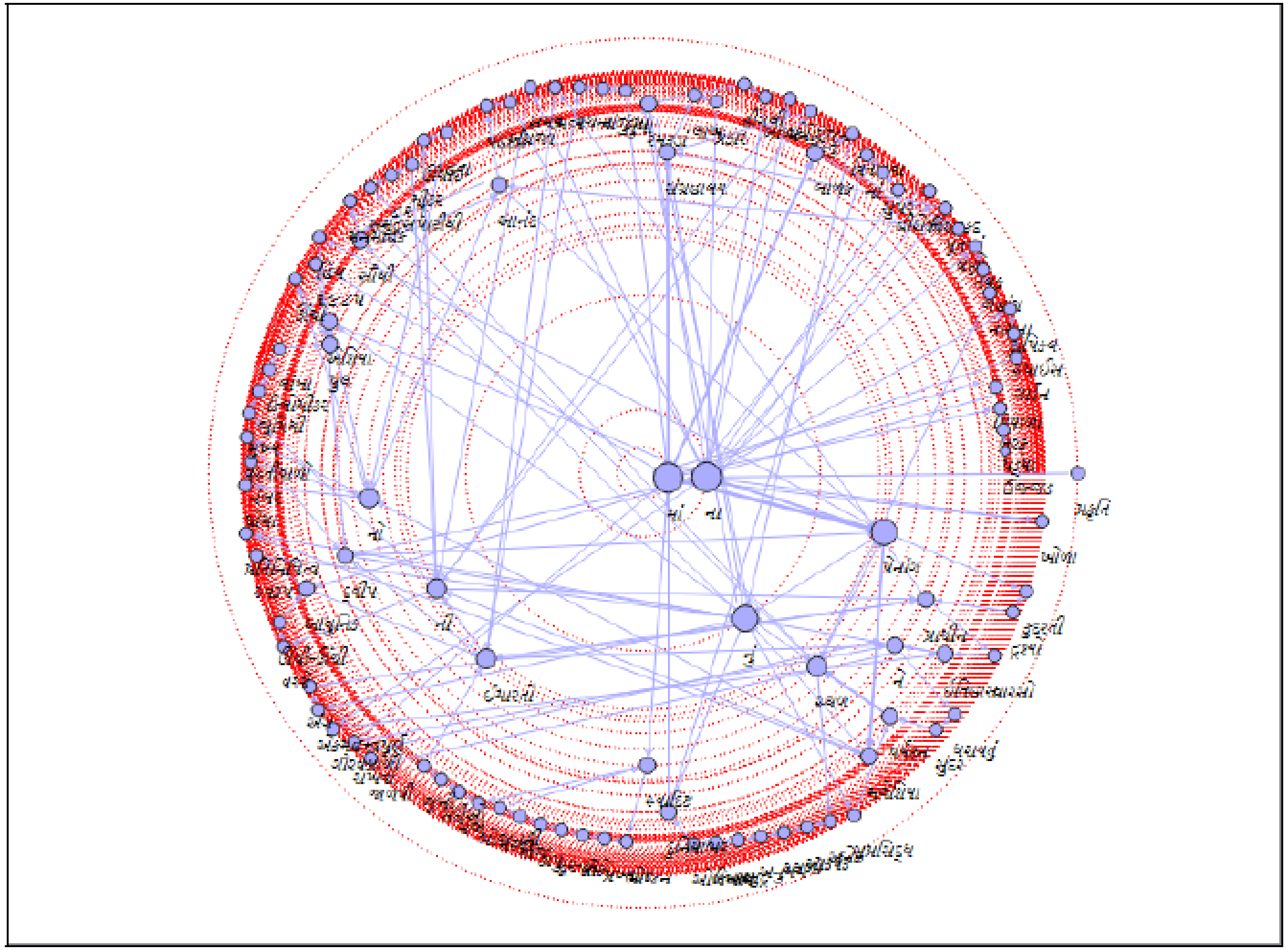


પ્રકૃતિના ખોળામાં આવેલું પેનાંગ કુદરતી દ્રશ્યો પ્રાચીન વારસો અને ઇતિહાસને ધરાવતું સુંદર પર્યટન સ્થળ છે પેનાંગ મલેશિયાનું એક સુપ્રસિદ્ધ પર્યટન સ્થળ છે તે પોતાના અનેક આકર્ષક સ્થળો તથા અનુપમ સમુદ્ર કિનારાઓ માટે તો ઓળખાય જ છે સાથે દુનિયાભરમાં પોતાના સ્વાદિષ્ટ ભોજન માટે પણ પ્રખ્યાત છે આ જ કારણ છે કે લોકો દૂર દૂરથી આ ટાપુ પર મોજ મસ્તી કરવા તથા અહીંના સ્વાદિષ્ટ ભોજનની મજા લેવા આવે છે અહીં આજે પણ સદીઓ જૂની ઈમારતોને જાળવીને રાખવામાં આવી છે જે પેનાંગના ગૌરવશાળી ઇતિહાસનું એકમહત્ત્વપૂર્ણ અંગ છે આ પ્રાચીન ઈમારતોની વચ્ચે તમને ઊંચી-ઊંચી આધુનિક ઈમારતો પણજોવા મળી જશે જે આ દ્વીપના આધુનિક સ્વરૂપનું પ્રતિનિધિત્વ કરે છે આ મલેશિયાનો ચોથા નંબરનો સૌથી મોટો તથા સૌથી વધારે વસ્તીવાળો દ્વીપ છે પેનાંગ મલેશિયાની મુખ્ય ભૂમિથી ૧૩.૫ કિલોમીટર લાંબા પુલ દ્વારા જોડાયેલું છે એશિયાનો આ સૌથી ઊંચો પુલ ૧૯૮૫માં બનીને તૈયાર થયો હતો પેનાંગ હિલ આ દ્વીપનું સૌથી મનમોહક સ્થળ છે જે સમુદ્રસપાટીથી ૮૨૧ મીટરની ઊંચાઈ પર આવેલ છે જ્યારે અમે તાજી હવાનો આનંદ લેતા આ પહાડી પર ચઢતાં જતાં હતાં ત્યારે અમને અંગ્રેજોના સમયની કેટલીક પ્રાચીન ભવ્ય ઈમારતો જોવા મળી જેની રચના તેમણે પોતાના માટે કરાવી હતી તાંજુંગા બુંગામાં આવેલ પેનાંગનું રમકડાંનું સંગ્રહાલય છે અહીં તમે ૧ લાખથી પણ વધારે જુદાં જુદાં પ્રકારના રમકડાં ઢીંગલીઓ તથા બાળકોના મનોરંજનની બીજી વસ્તુઓ જોઈ શકો છો આ સંગ્રહાલયમાં બાળકો વધારે સમય વિતાવવાનું પસંદ કરે છે અહીં દુનિયાભરનાં સુપર હીરોઝના વિરાટકદના પૂતળાં પણ મૂકવામાં આવ્યા છે આ સંગ્રહાલયમાં અમારી સાથે આવેલા બાળકોએ ઘણો આનંદ કર્યો અમને સૌથી વધારે આનંદ પેનાંગના તેલુક બહાંગ નામના સ્થળે આવેલ એશિયાના એકમાત્ર ટ્રોપિકલ સ્પાઈસ ગાર્ડનમાં આવ્યો ભરેલી રહે છે પરંતુ શિયાળામાં બરફ પડ્યા પછી ઉજ્જડ બની જાય છે.

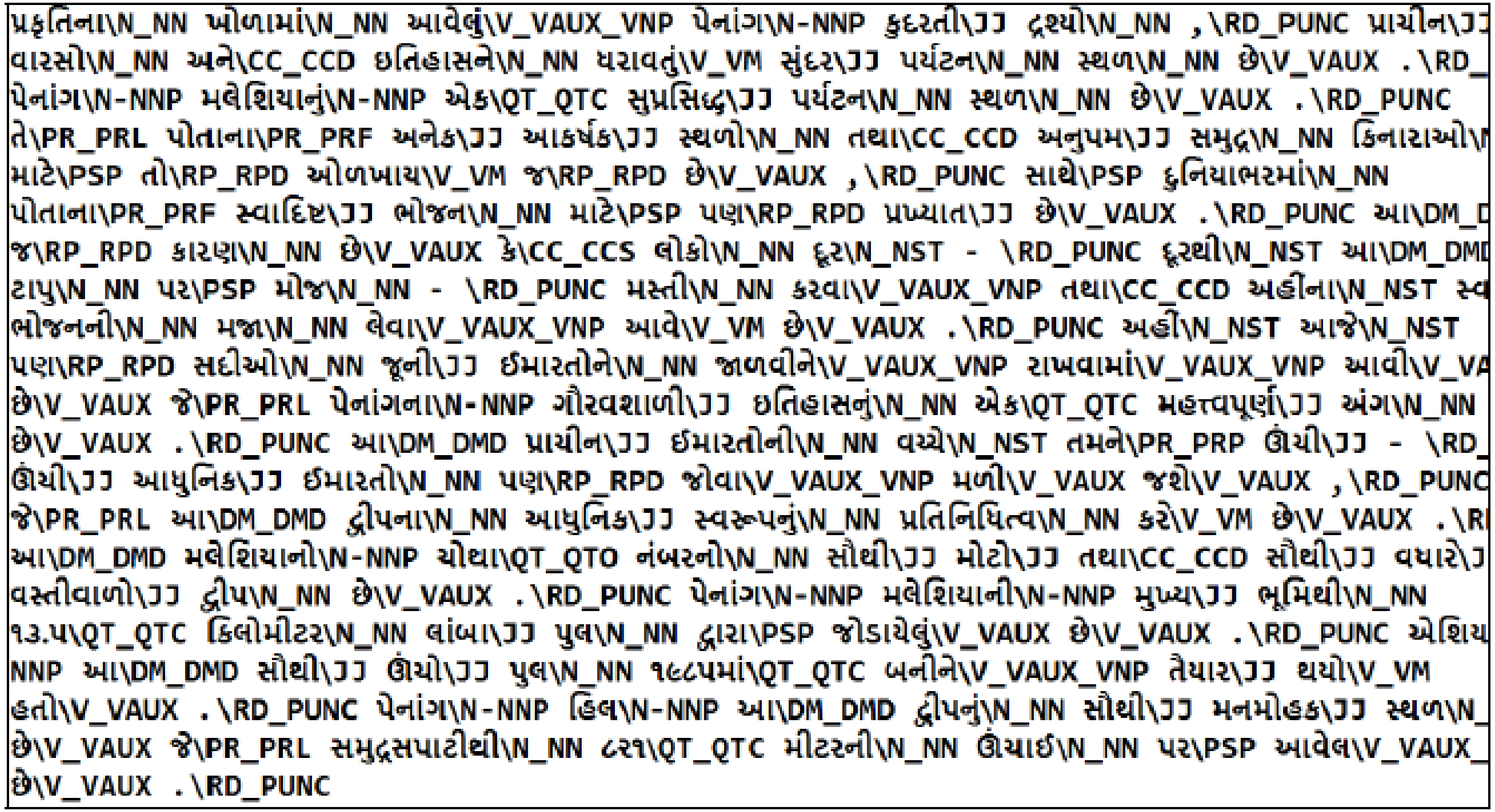
Graph from Text Document



Finding the Significant Terms from Text

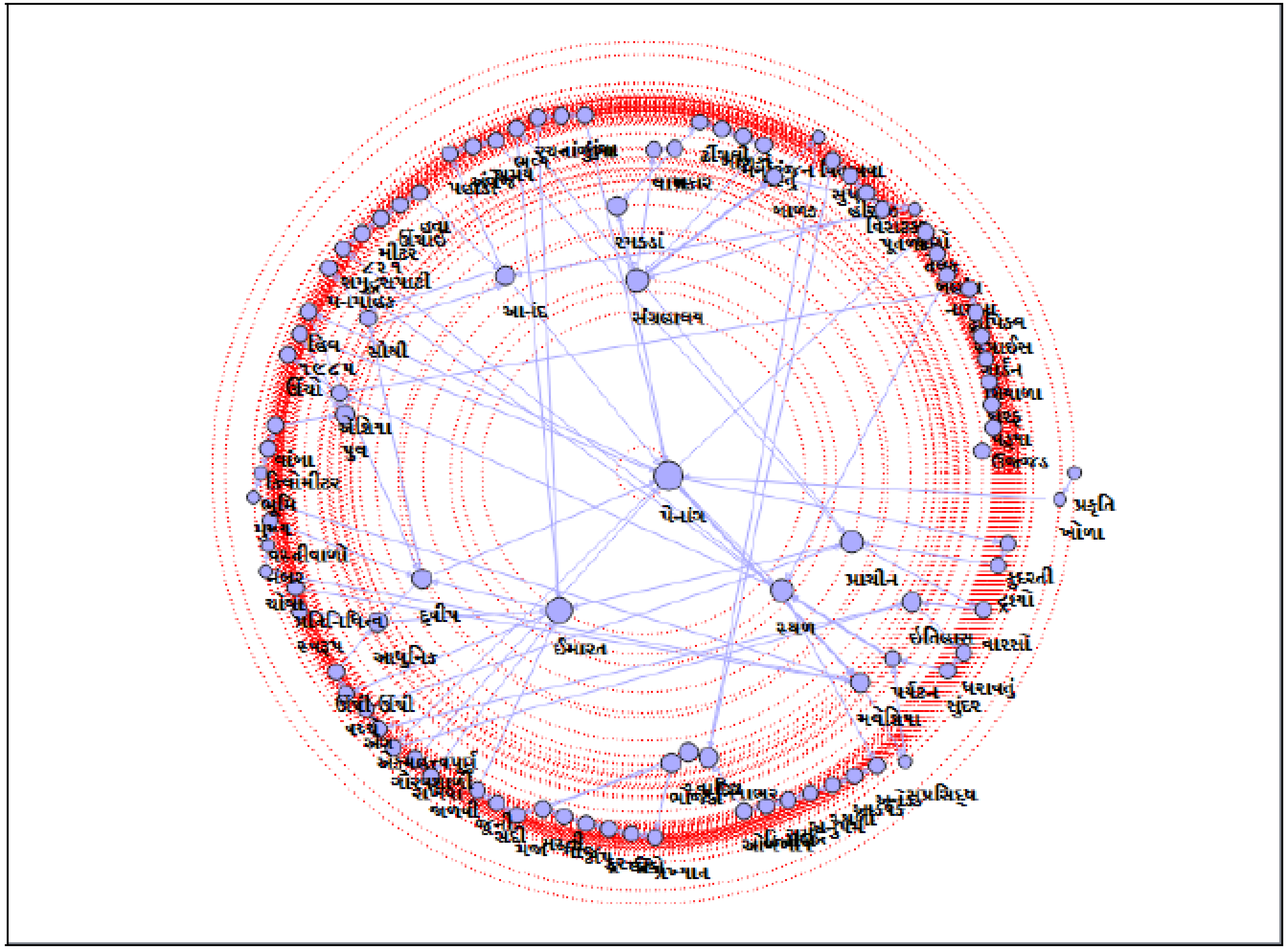


Processed Text – POS Tagging

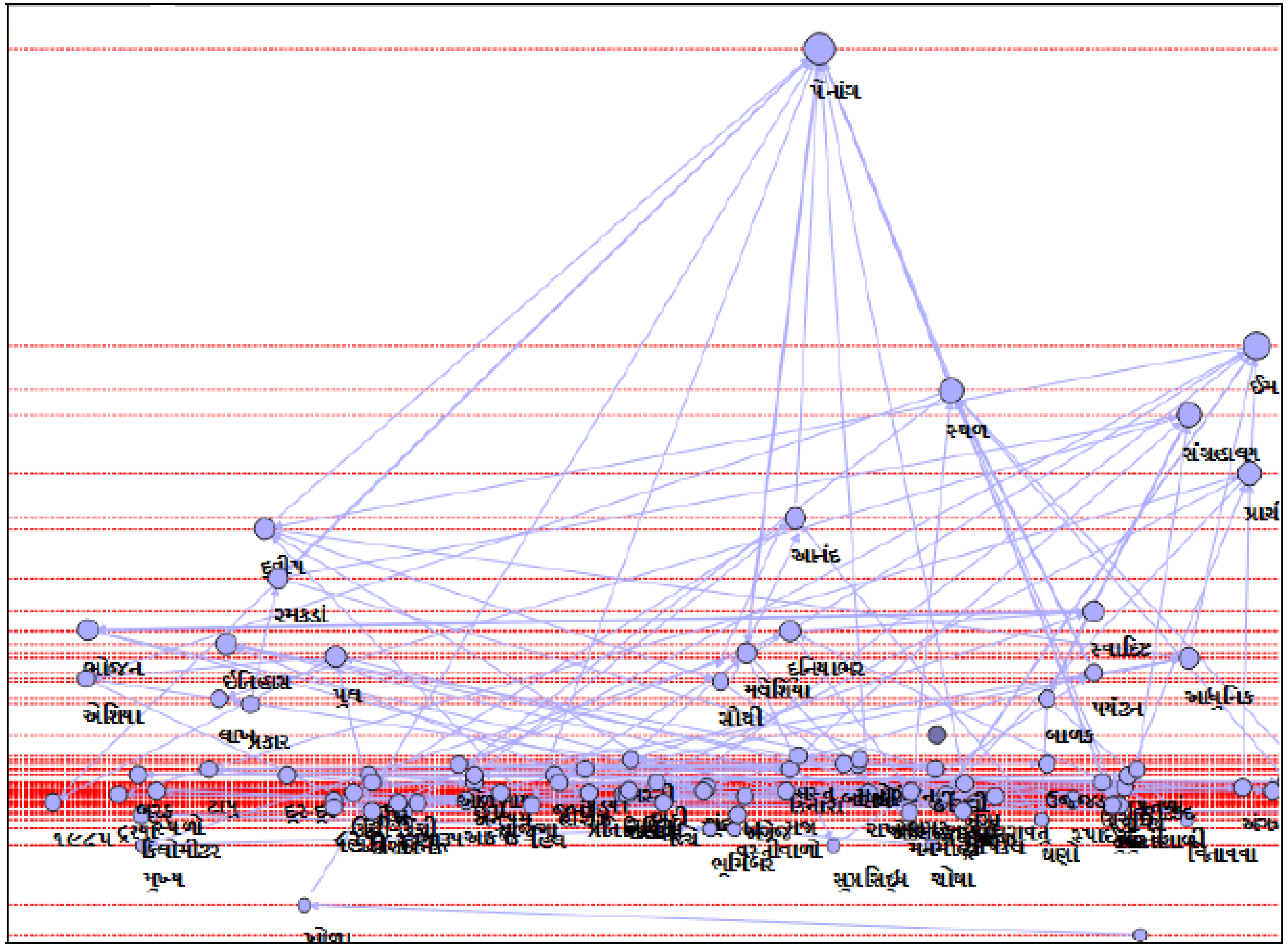


```
પ્રકૃતિના\N_NN ખોળામાં\N_NN આવેલું\V_VAUX_VNP પેનાંગ\N-NNP કુદરતી\JJ દ્રશ્યો\N_NN ,\RD_PUNC પ્રાચીન\JJ
વારસો\N_NN અને\CC_CCD ઇતિહાસને\N_NN ધરાવતું\V_VM સુંદર\JJ પર્યટન\N_NN સ્થળ\N_NN છે\V_VAUX .\RD_
પેનાંગ\N-NNP મલેશિયાનું\N-NNP એક\QT_QTC સુપ્રસિદ્ધ\JJ પર્યટન\N_NN સ્થળ\N_NN છે\V_VAUX .\RD_PUNC
તે\PR_PRL પોતાના\PR_PRF અનેક\JJ આકર્ષક\JJ સ્થળો\N_NN તથા\CC_CCD અનુપમ\JJ સમુદ્ર\N_NN કિનારાઓ\N
માટે\PSP તો\RP_RPD ઓળખાય\V_VM જ\RP_RPD છે\V_VAUX ,\RD_PUNC સાથે\PSP દુનિયાભરમાં\N_NN
પોતાના\PR_PRF સ્વાદિષ્ટ\JJ ભોજન\N_NN માટે\PSP પણ\RP_RPD પ્રખ્યાત\JJ છે\V_VAUX .\RD_PUNC આ\DM_D
જ\RP_RPD કારણ\N_NN છે\V_VAUX કે\CC_CCS લોકો\N_NN દૂર\N_NST - \RD_PUNC દૂરથી\N_NST આ\DM_DMD
ટાપુ\N_NN પર\PSP મોજ\N_NN - \RD_PUNC મસ્તી\N_NN કરવા\V_VAUX_VNP તથા\CC_CCD અહીંના\N_NST સ્વ
ભોજનની\N_NN મજા\N_NN લેવા\V_VAUX_VNP આવે\V_VM છે\V_VAUX .\RD_PUNC અહીં\N_NST આજે\N_NST
પણ\RP_RPD સદીઓ\N_NN જૂની\JJ ઈમારતોને\N_NN જાળવીને\V_VAUX_VNP રાખવામાં\V_VAUX_VNP આવી\V_VA
છે\V_VAUX જે\PR_PRL પેનાંગના\N-NNP ગૌરવશાળી\JJ ઇતિહાસનું\N_NN એક\QT_QTC મહત્વપૂર્ણ\JJ અંગ\N_NN
છે\V_VAUX .\RD_PUNC આ\DM_DMD પ્રાચીન\JJ ઈમારતોની\N_NN વચ્ચે\N_NST તમને\PR_PRP ઊંચી\JJ - \RD_
ઊંચી\JJ આધુનિક\JJ ઈમારતો\N_NN પણ\RP_RPD જોવા\V_VAUX_VNP મળી\V_VAUX જશે\V_VAUX ,\RD_PUNC
જે\PR_PRL આ\DM_DMD દ્વીપના\N_NN આધુનિક\JJ સ્વરૂપનું\N_NN પ્રતિનિધિત્વ\N_NN કરે\V_VM છે\V_VAUX .\R
આ\DM_DMD મલેશિયાનો\N-NNP ચોથા\QT_QTO નંબરનો\N_NN સૌથી\JJ મોટો\JJ તથા\CC_CCD સૌથી\JJ વધારે\J
વસ્તીવાળો\JJ દ્વીપ\N_NN છે\V_VAUX .\RD_PUNC પેનાંગ\N-NNP મલેશિયાની\N-NNP મુખ્ય\JJ ભૂમિથી\N_NN
૧૩.૫\QT_QTC કિલોમીટર\N_NN લાંબા\JJ પુલ\N_NN દ્વારા\PSP જોડાયેલું\V_VAUX છે\V_VAUX .\RD_PUNC એશિય
NNP આ\DM_DMD સૌથી\JJ ઊંચો\JJ પુલ\N_NN ૧૯૮૫માં\QT_QTC બનીને\V_VAUX_VNP તૈયાર\JJ થયો\V_VM
હતો\V_VAUX .\RD_PUNC પેનાંગ\N-NNP હિલ\N-NNP આ\DM_DMD દ્વીપનું\N_NN સૌથી\JJ મનમોહક\JJ સ્થળ\N_
છે\V_VAUX જે\PR_PRL સમુદ્રસપાટીથી\N_NN ૮૨૧\QT_QTC મીટરની\N_NN ઊંચાઈ\N_NN પર\PSP આવેલ\V_VAUX_
છે\V_VAUX .\RD_PUNC
```

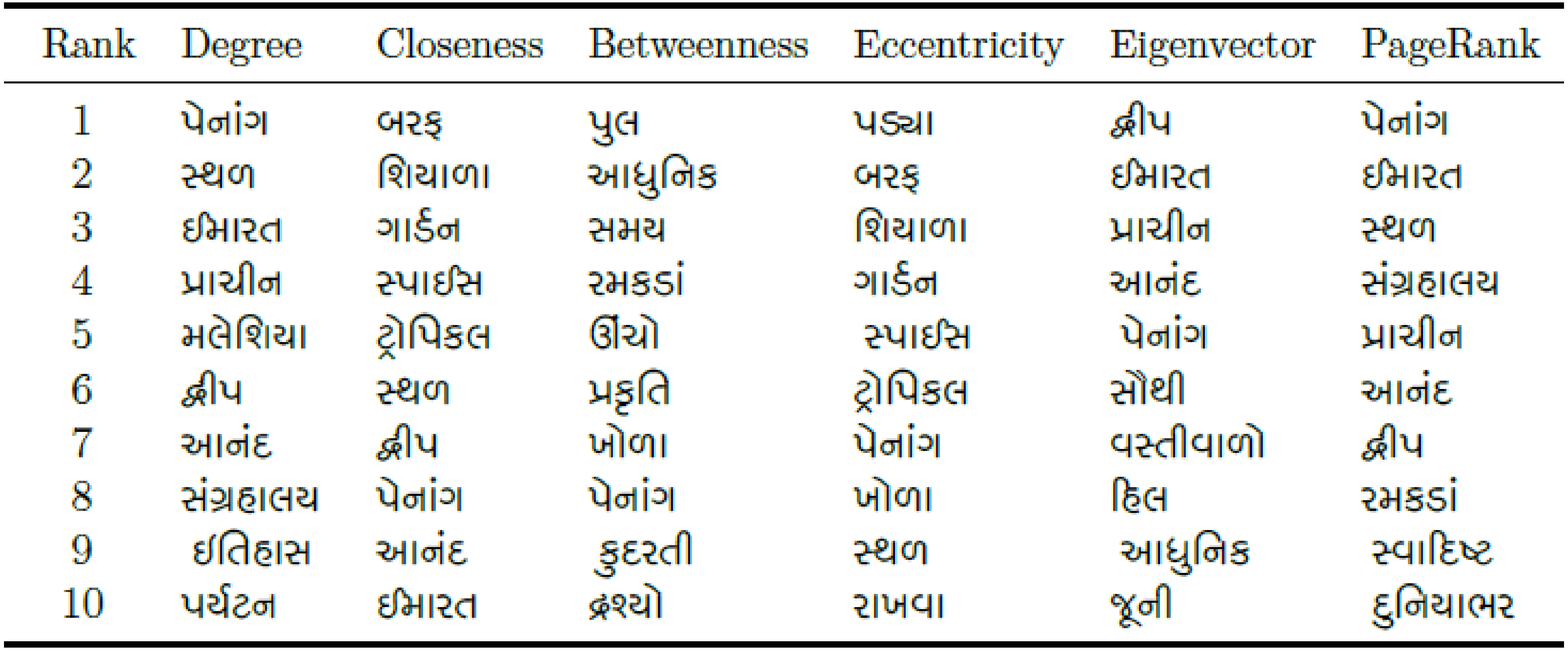
Graph from Processed Text



Graph from Processed Text

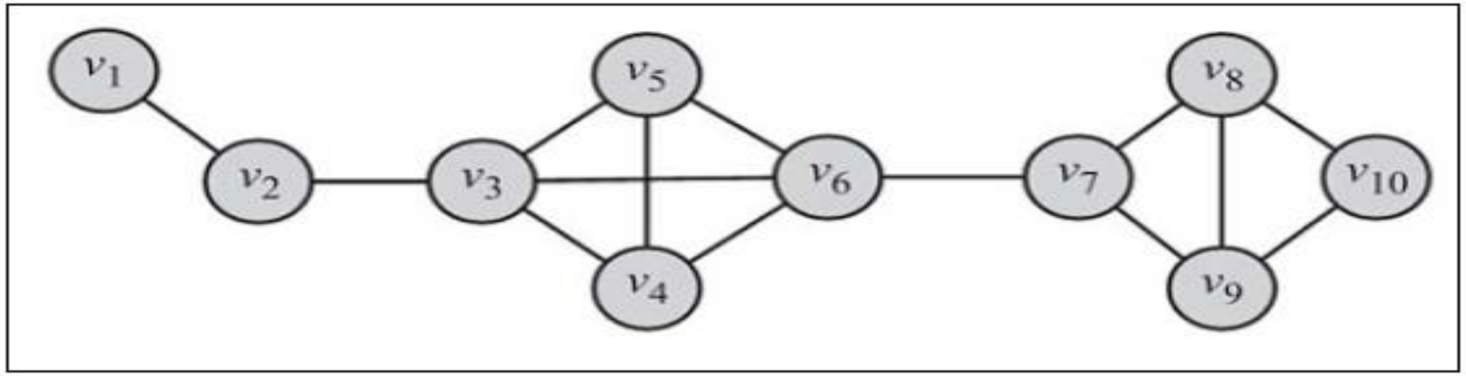


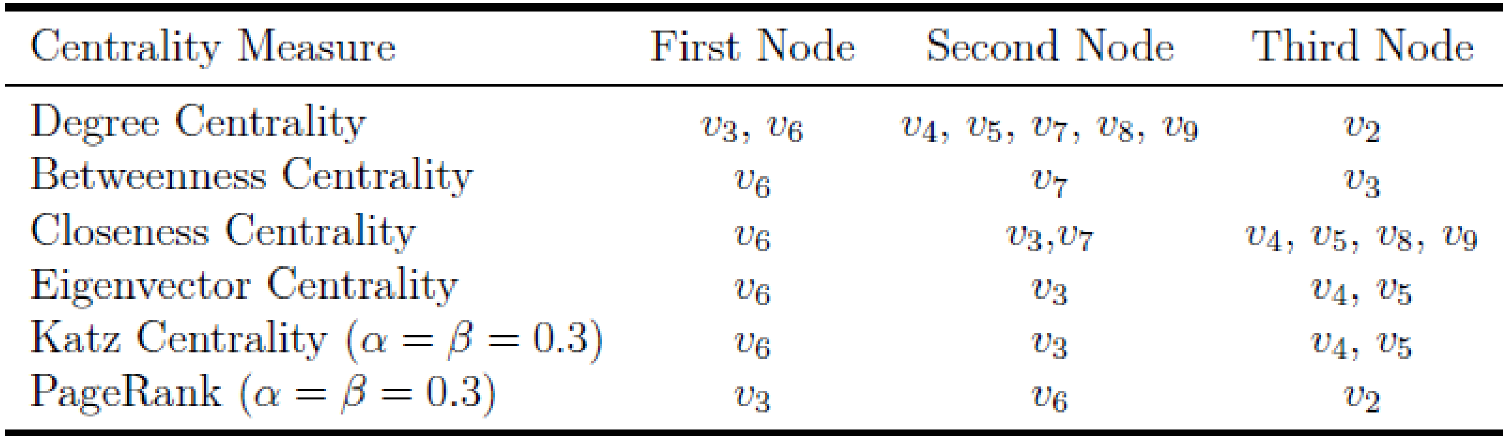
Significant Terms from the Text



Rank	Degree	Closeness	Betweenness	Eccentricity	Eigenvector	PageRank
1	પેનાંગ	બરફ	પુલ	પડ્યા	દ્વીપ	પેનાંગ
2	સ્થળ	શિયાળા	આધુનિક	બરફ	ઇમારત	ઇમારત
3	ઇમારત	ગાર્ડન	સમય	શિયાળા	પ્રાચીન	સ્થળ
4	પ્રાચીન	સ્પાઇસ	રમકડાં	ગાર્ડન	આનંદ	સંગ્રહાલય
5	મલેશિયા	ટ્રોપિકલ	ઊંચો	સ્પાઇસ	પેનાંગ	પ્રાચીન
6	દ્વીપ	સ્થળ	પ્રકૃતિ	ટ્રોપિકલ	સૌથી	આનંદ
7	આનંદ	દ્વીપ	ખોળા	પેનાંગ	વસ્તીવાળો	દ્વીપ
8	સંગ્રહાલય	પેનાંગ	પેનાંગ	ખોળા	હિલ	રમકડાં
9	ઇતિહાસ	આનંદ	કુદરતી	સ્થળ	આધુનિક	સ્વાદિષ્ટ
10	પર્યટન	ઇમારત	દ્રશ્યો	રાખવા	જૂની	દુનિયાભર

Centrality Measures for Graphs

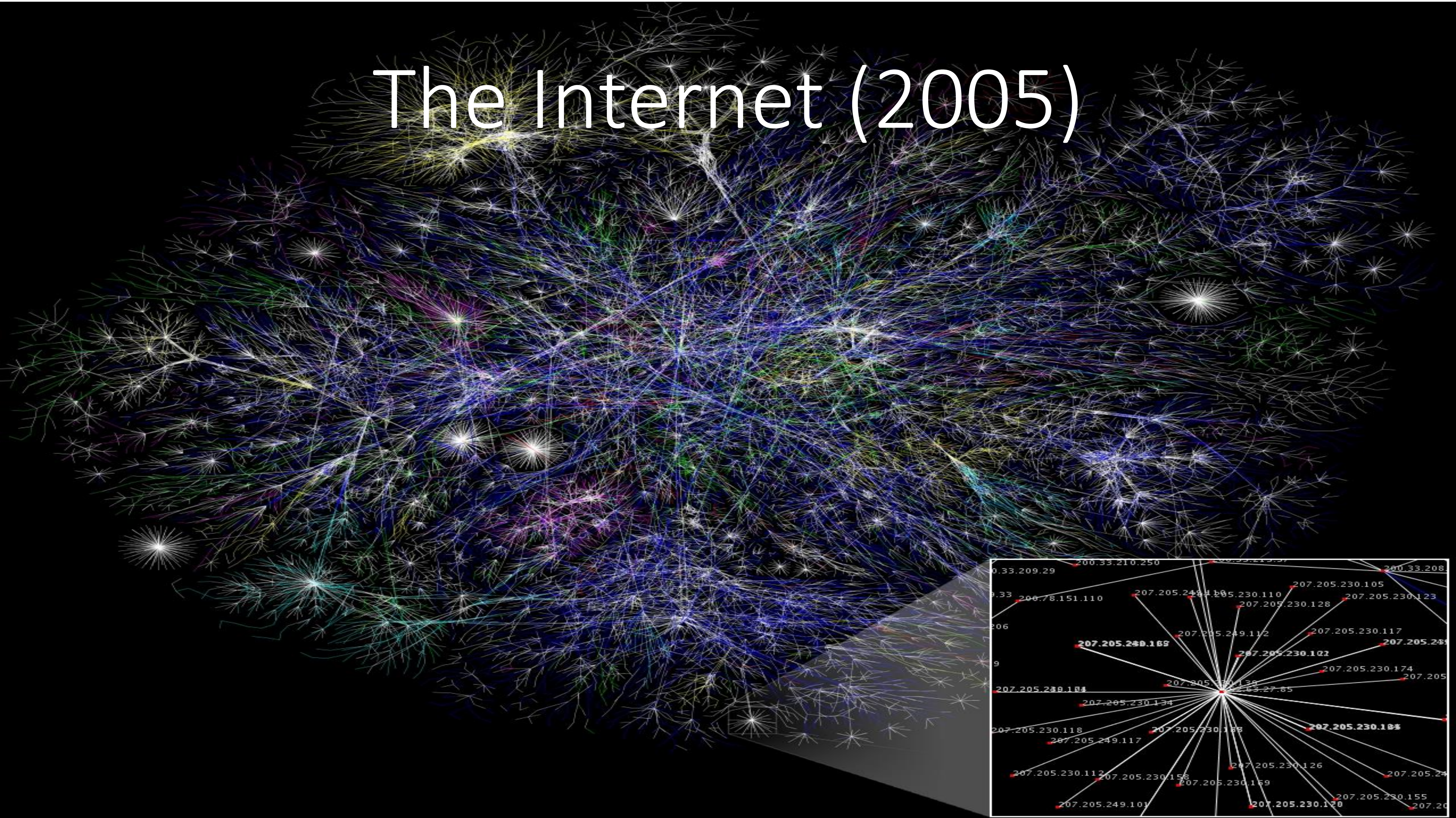




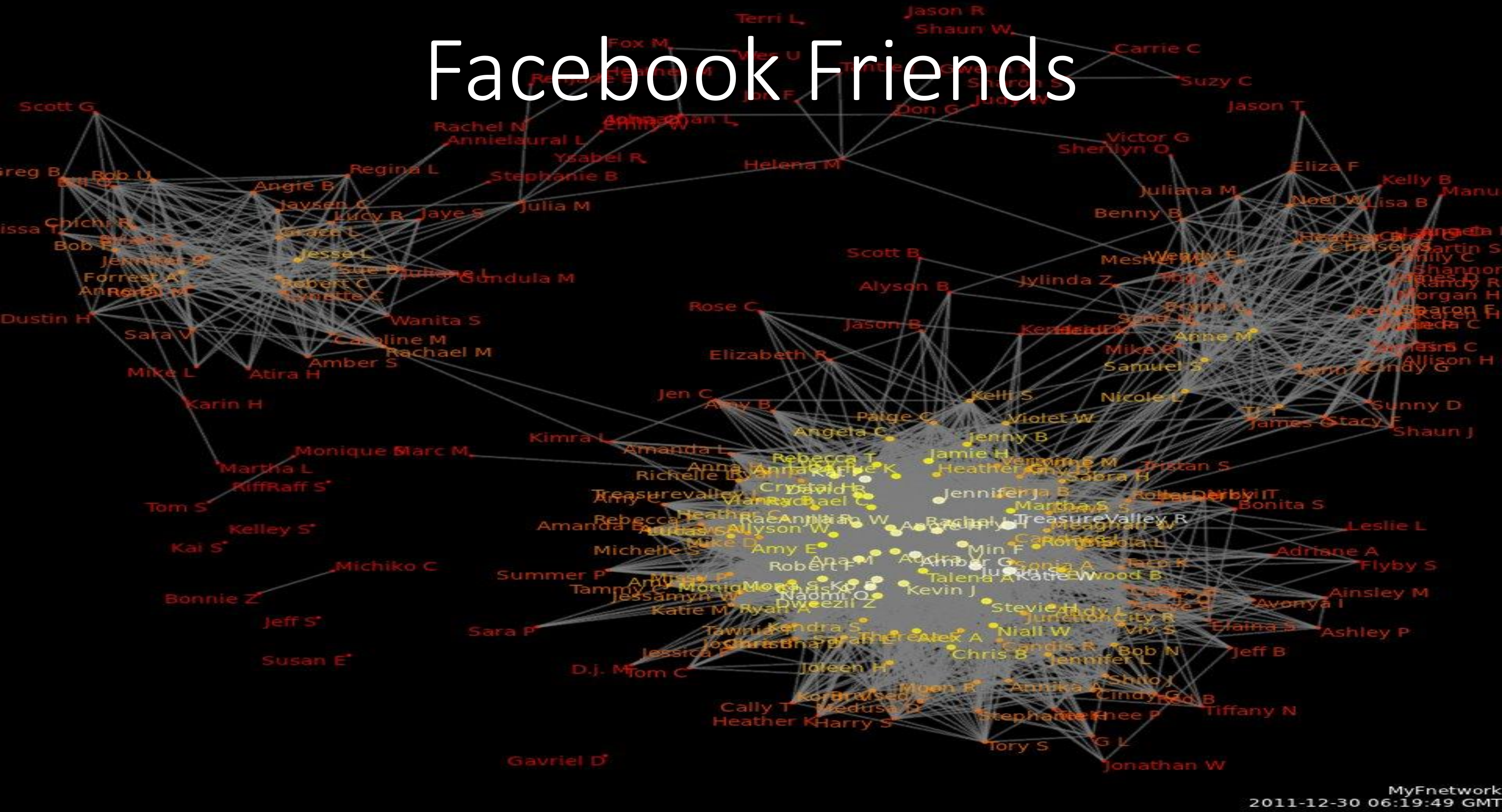
Centrality Measure	First Node	Second Node	Third Node
Degree Centrality	v_3, v_6	v_4, v_5, v_7, v_8, v_9	v_2
Betweenness Centrality	v_6	v_7	v_3
Closeness Centrality	v_6	v_3,v_7	v_4, v_5, v_8, v_9
Eigenvector Centrality	v_6	v_3	v_4, v_5
Katz Centrality ($\alpha = \beta = 0.3$)	v_6	v_3	v_4, v_5
PageRank ($\alpha = \beta = 0.3$)	v_3	v_6	v_2

Dealing with Graphs

The Internet (2005)



Facebook Friends



Deriving Meaning from Graphs

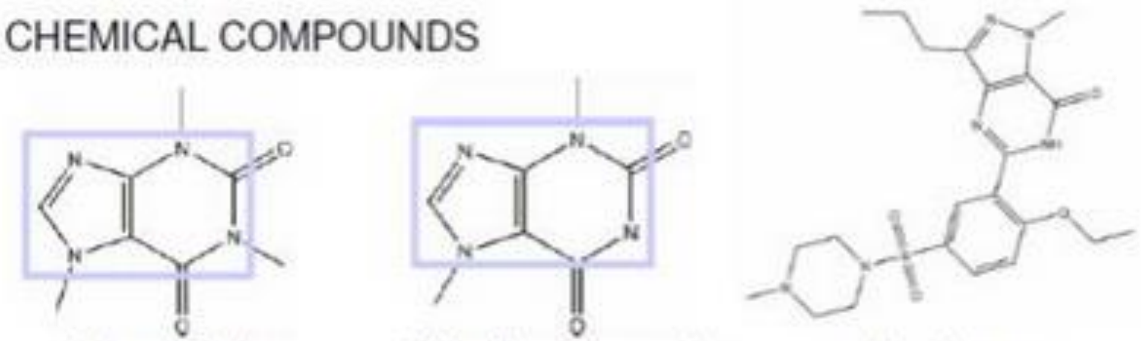
Graph analytics is commonly used term, and it refers specifically to the process of analyzing data in a **graph** format using data points as nodes and relationships as edges.

Graph Mining is the set of tools and techniques used to (a) analyze the properties of real-world graphs, (b) predict how the structure and properties of a given graph might affect some application, and (c) develop models that can generate realistic graphs that match the patterns found in real-world graphs of interest.

https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_350

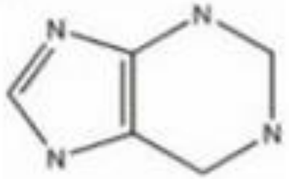
Graph Mining in Chemical Compounds

CHEMICAL COMPOUNDS



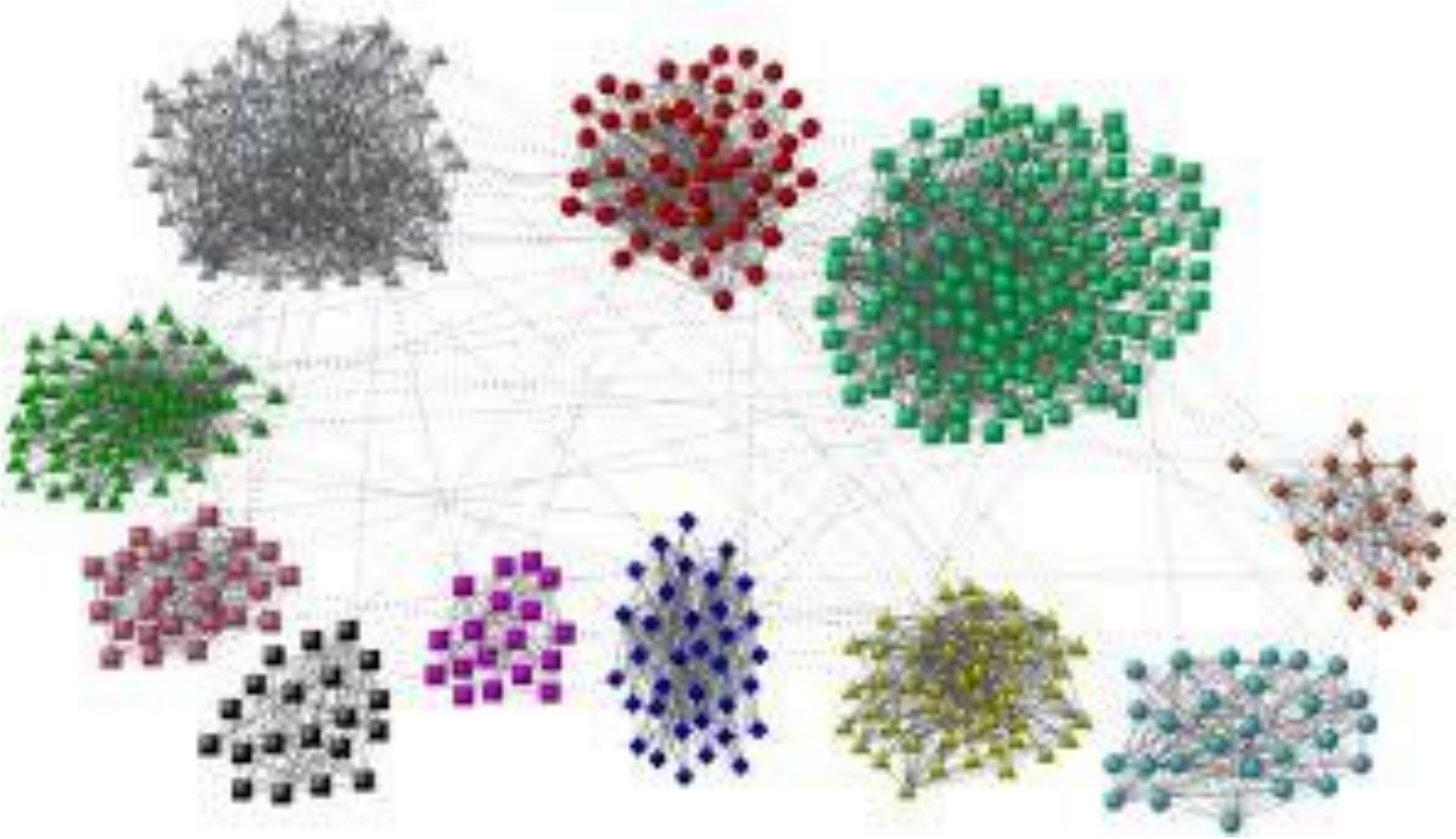
(a) caffeine (b) diurobromine (c) viagra …

FREQUENT SUBGRAPH

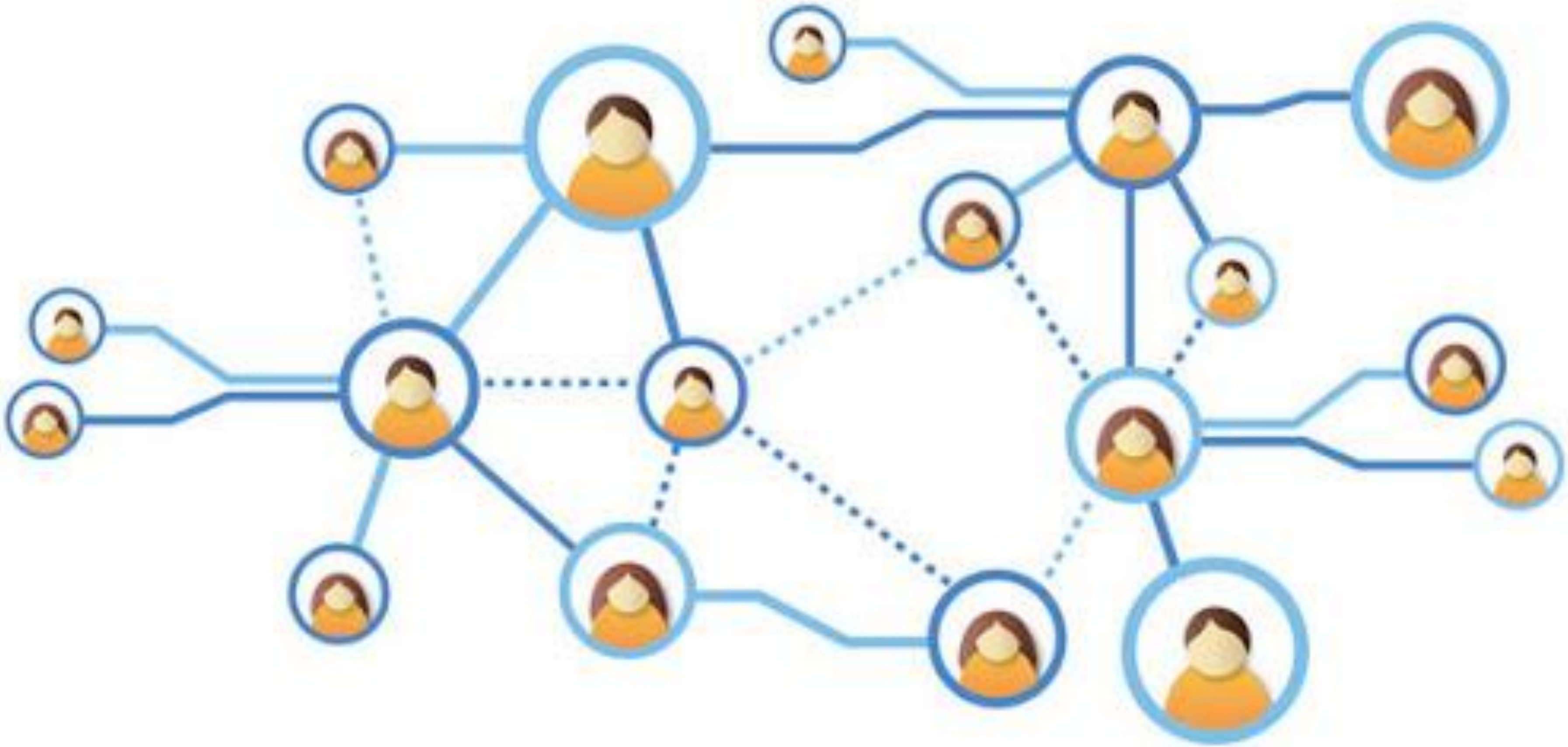


'From K. Borgwardt and X. Yan (KDD'08)

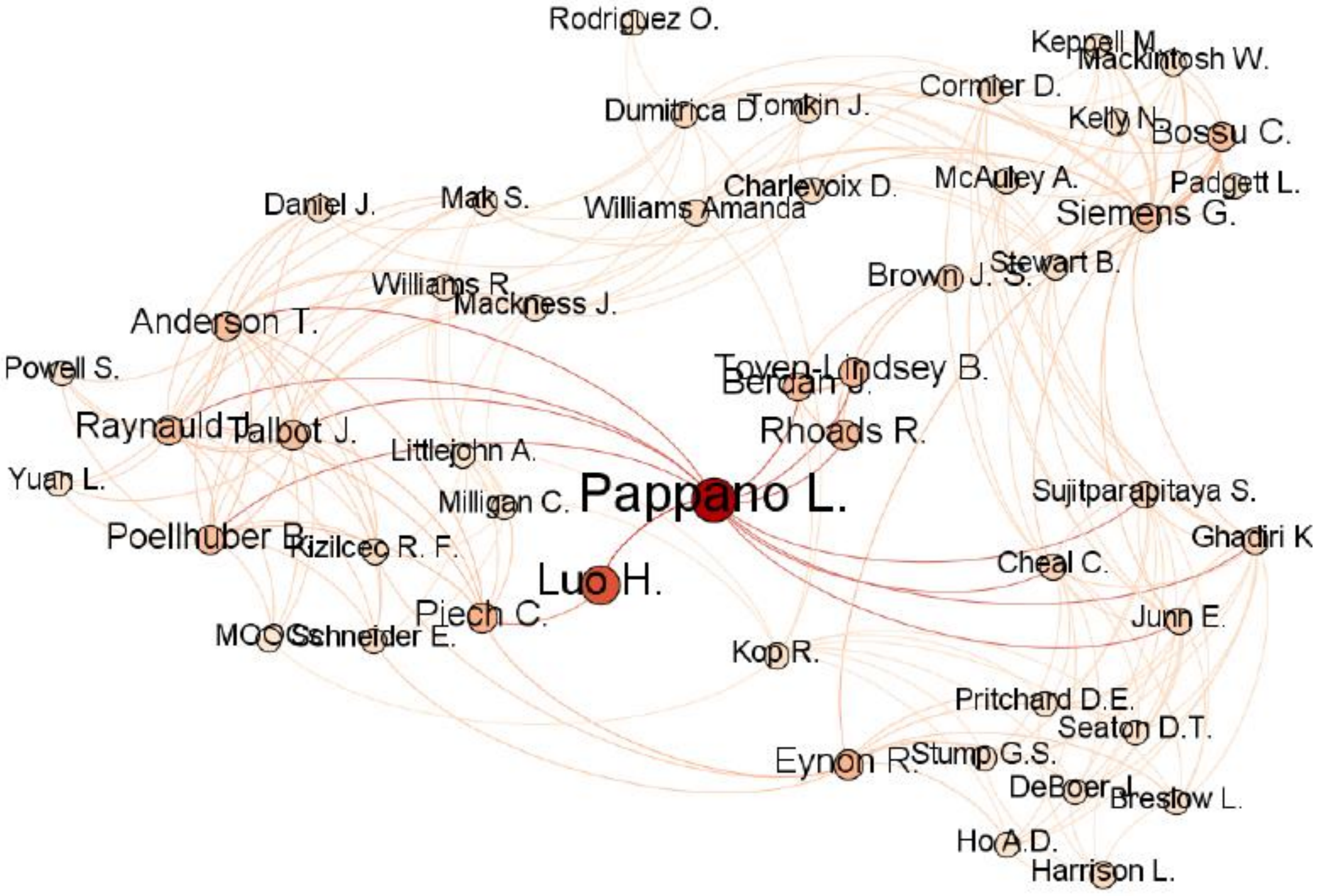
Community Detection



Finding Influential Persons



Citation Network



Few Questions that can be answered by Graph Analytics

- How to travel (best path as per different scenario) from one person to other
- The longest of all shortest paths
- The largest distance between given node and all other nodes
- Determine closeness to all other nodes.
- Which person can convey information to many other persons
- Which person stand between groups is network
- How many communities exist
- Which all persons making close groups
- How they are forming natural group while being similar or dissimilar.
- The natural group of various people from different dimensions

Graph Mining Applications

- Pandemic Situation (Spreading of infection)
- Web Graphs (Pages & Hyperlinks)
- Social Science Graphs (Social Media & Friends)
- Computer Networks Graphs (Routers, Network Traffic)
- Computer Security (Behavior of malwares, spread, intruders)
- Biological Graphs (Biomolecules, Neurons, Transport Systems)
- Chemical Graphs (Chemical Structures, DNA)
- Finance Graphs

Graph Mining Applications

- Healthcare Graphs (Doctors, Lawyers & Claims)
- Software Engineering Graphs (Operations & Dependencies)
- Climatology
- Entertainment (Movies, Actors, Genre, Awards)
- Research (Citations, Co-authors)
- Crime (Finger Print matching)
- Transportation Data (Airlines, Railways Network, etc.)

Widely Used Social Network Analysis & Visualization Software

- Gephi - visualization
- Graphviz - visualization
- Igraph (Package) – creating & manipulating graphs
- JUNG (Java Universal Network Graph) library
- Mathematica
- NodeXL
- NetMiner
- Networkx (python library)

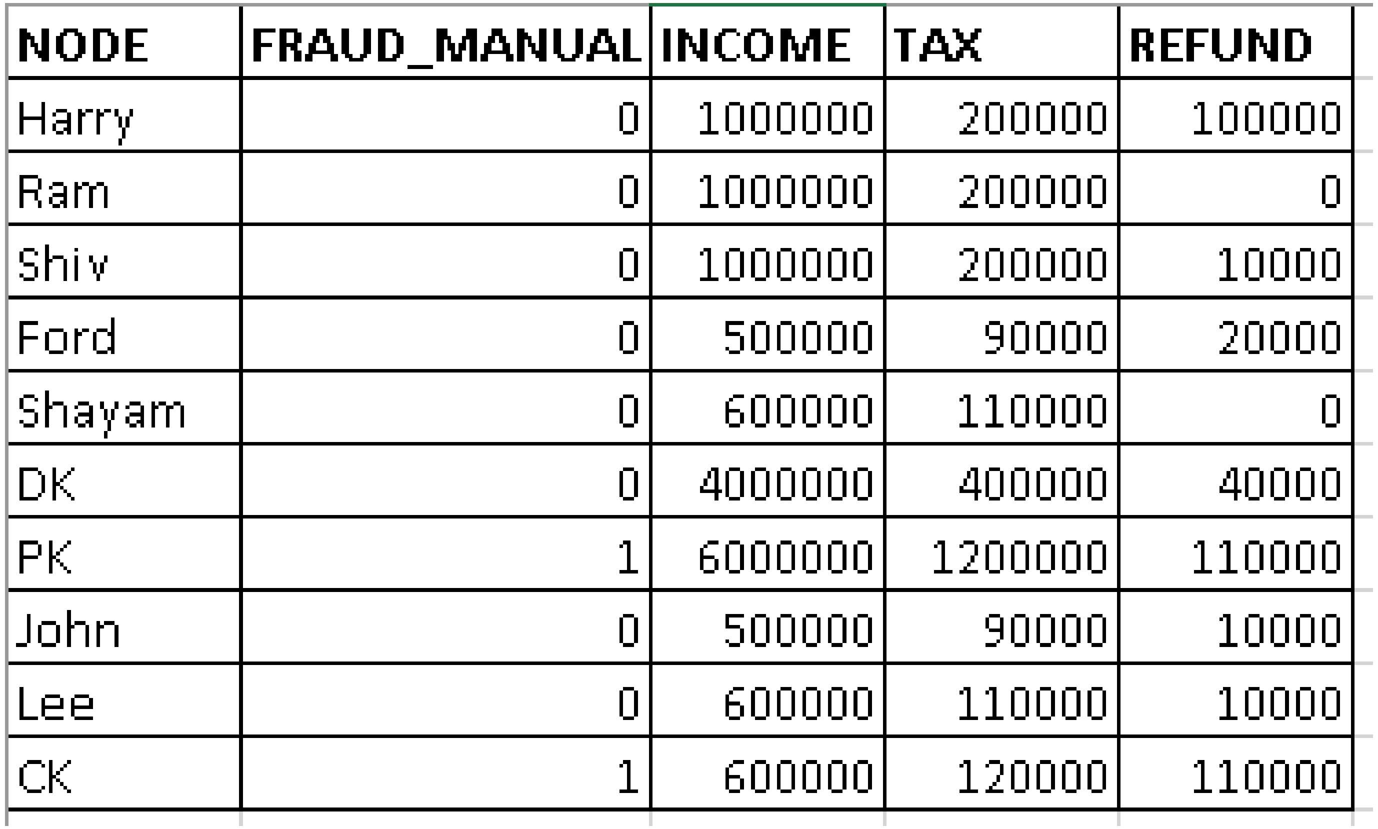
https://en.wikipedia.org/wiki/Social_network_analysis_software

Let us Practice …

- Python 3.7.4 version & Spyder IDE
- https://repo.anaconda.com/archive/Anaconda3-2019.10-Windows-x86_64.exe

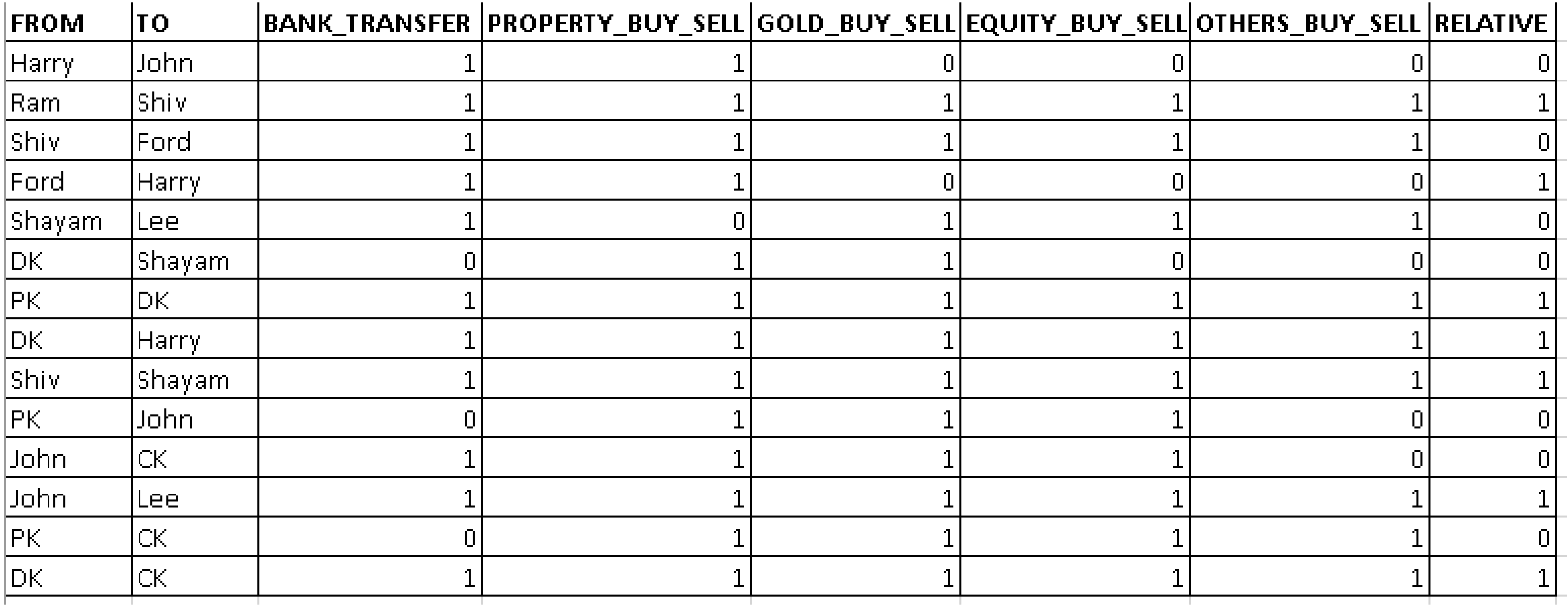
- pip install -q networkx
- pip install -q adjustText
- pip install -q nxviz
- pip install node2vec

Example – 1 (Nodes Data)



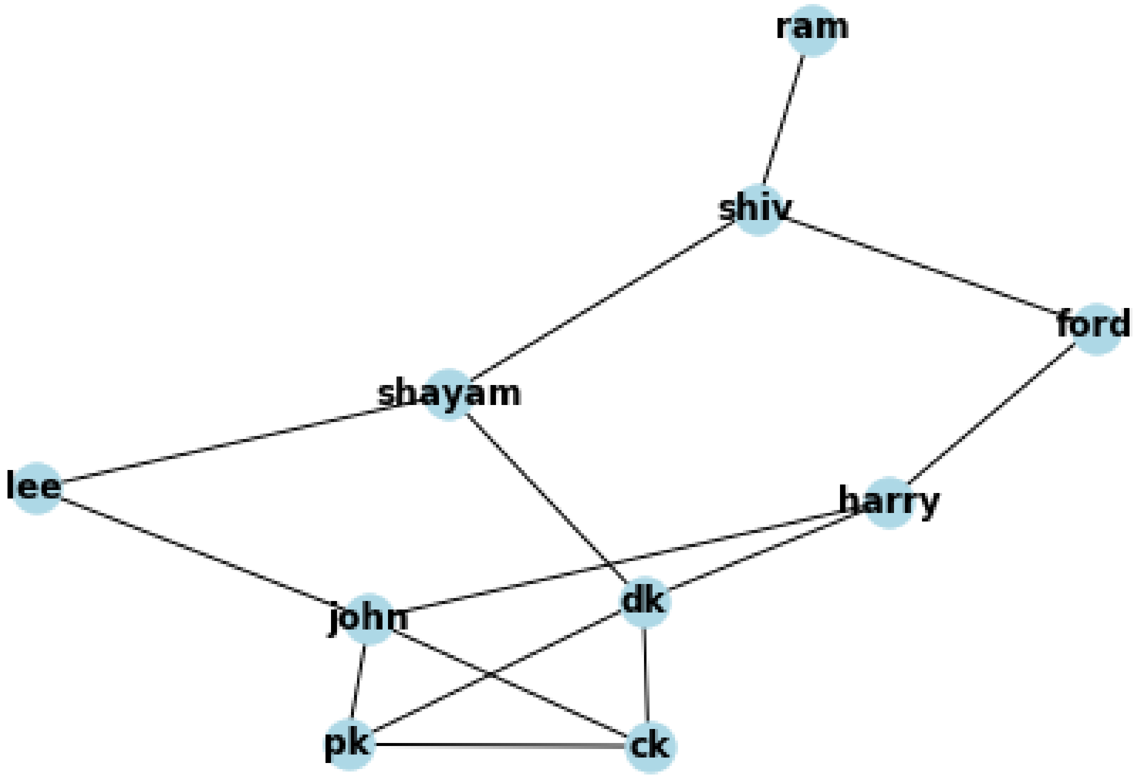
NODE	FRAUD_MANUAL	INCOME	TAX	REFUND
Harry	0	1000000	200000	100000
Ram	0	1000000	200000	0
Shiv	0	1000000	200000	10000
Ford	0	500000	90000	20000
Shayam	0	600000	110000	0
DK	0	4000000	400000	40000
PK	1	6000000	1200000	110000
John	0	500000	90000	10000
Lee	0	600000	110000	10000
CK	1	600000	120000	110000

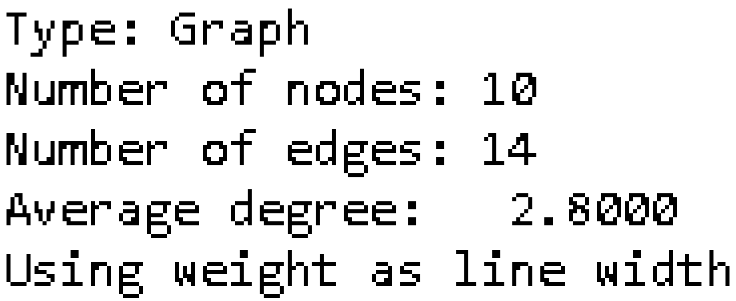
Example – 1 (Edge Data)



FROM	TO	BANK_TRANSFER	PROPERTY_BUY_SELL	GOLD_BUY_SELL	EQUITY_BUY_SELL	OTHERS_BUY_SELL	RELATIVE
Harry	John	1	1	0	0	0	0
Ram	Shiv	1	1	1	1	1	1
Shiv	Ford	1	1	1	1	1	0
Ford	Harry	1	1	0	0	0	1
Shayam	Lee	1	0	1	1	1	0
DK	Shayam	0	1	1	0	0	0
PK	DK	1	1	1	1	1	1
DK	Harry	1	1	1	1	1	1
Shiv	Shayam	1	1	1	1	1	1
PK	John	0	1	1	1	0	0
John	CK	1	1	1	1	0	0
John	Lee	1	1	1	1	1	1
PK	CK	0	1	1	1	1	0
DK	CK	1	1	1	1	1	1

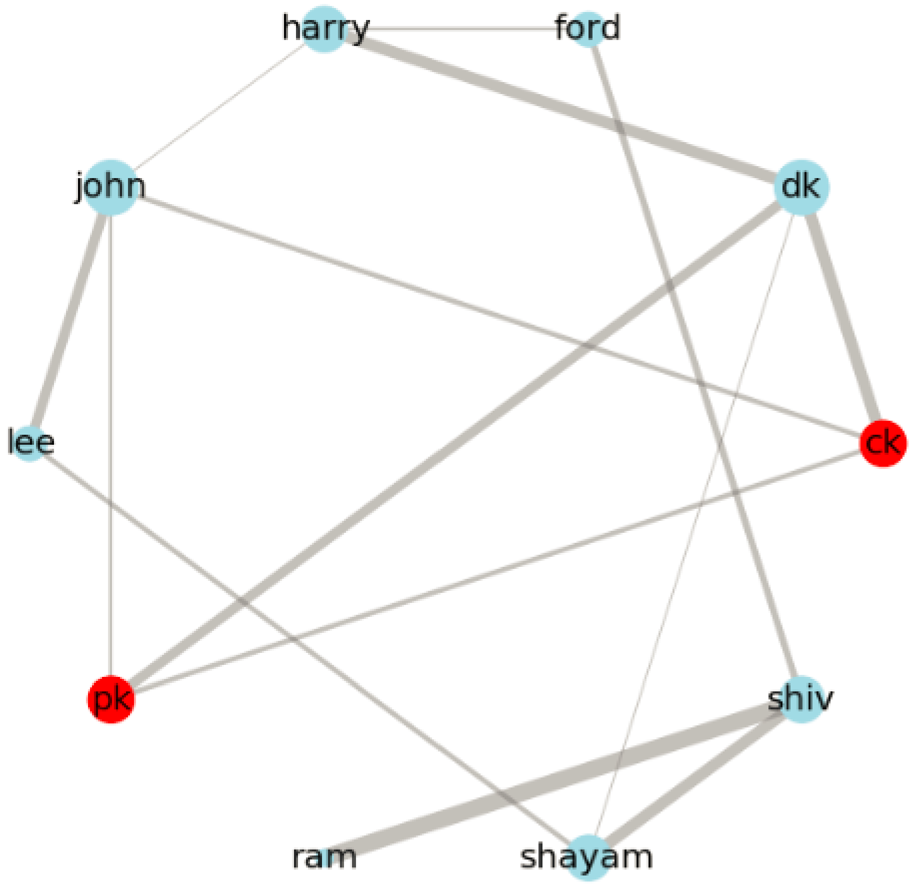
Graph using networkx package



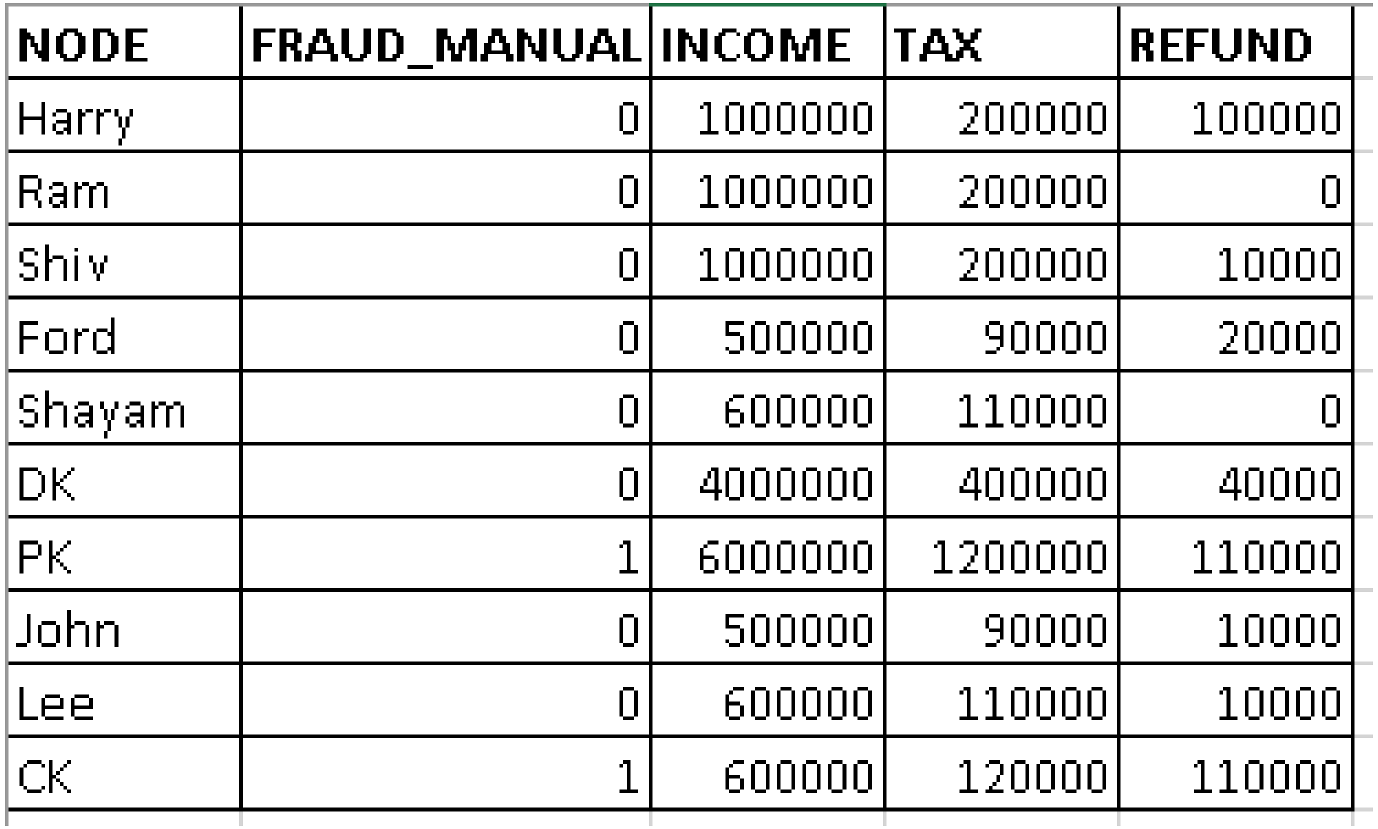


```
Type: Graph
Number of nodes: 10
Number of edges: 14
Average degree:   2.8000
Using weight as line width
```

Identifying Nodes & Edges (weights)

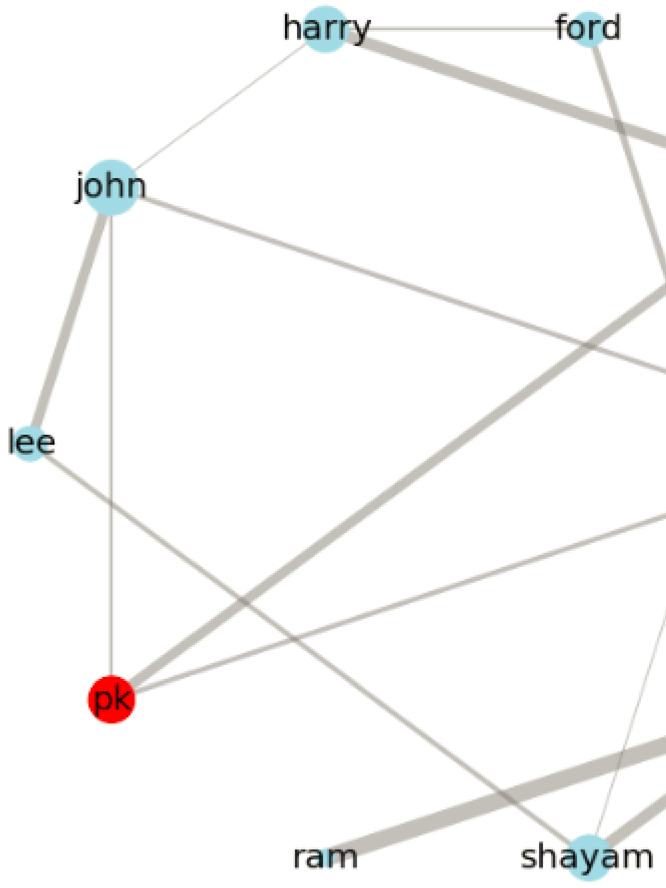


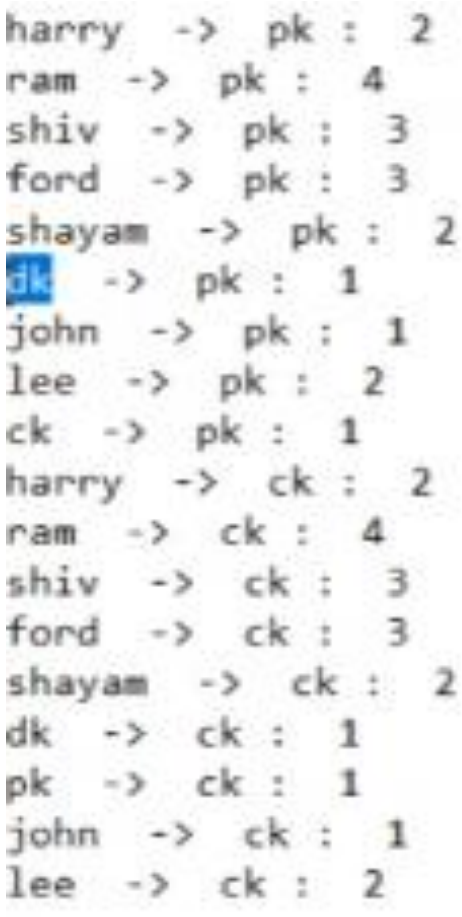
Example – 1 (Frauds)



NODE	FRAUD_MANUAL	INCOME	TAX	REFUND
Harry	0	1000000	200000	100000
Ram	0	1000000	200000	0
Shiv	0	1000000	200000	10000
Ford	0	500000	90000	20000
Shayam	0	600000	110000	0
DK	0	4000000	400000	40000
PK	1	6000000	1200000	110000
John	0	500000	90000	10000
Lee	0	600000	110000	10000
CK	1	600000	120000	110000

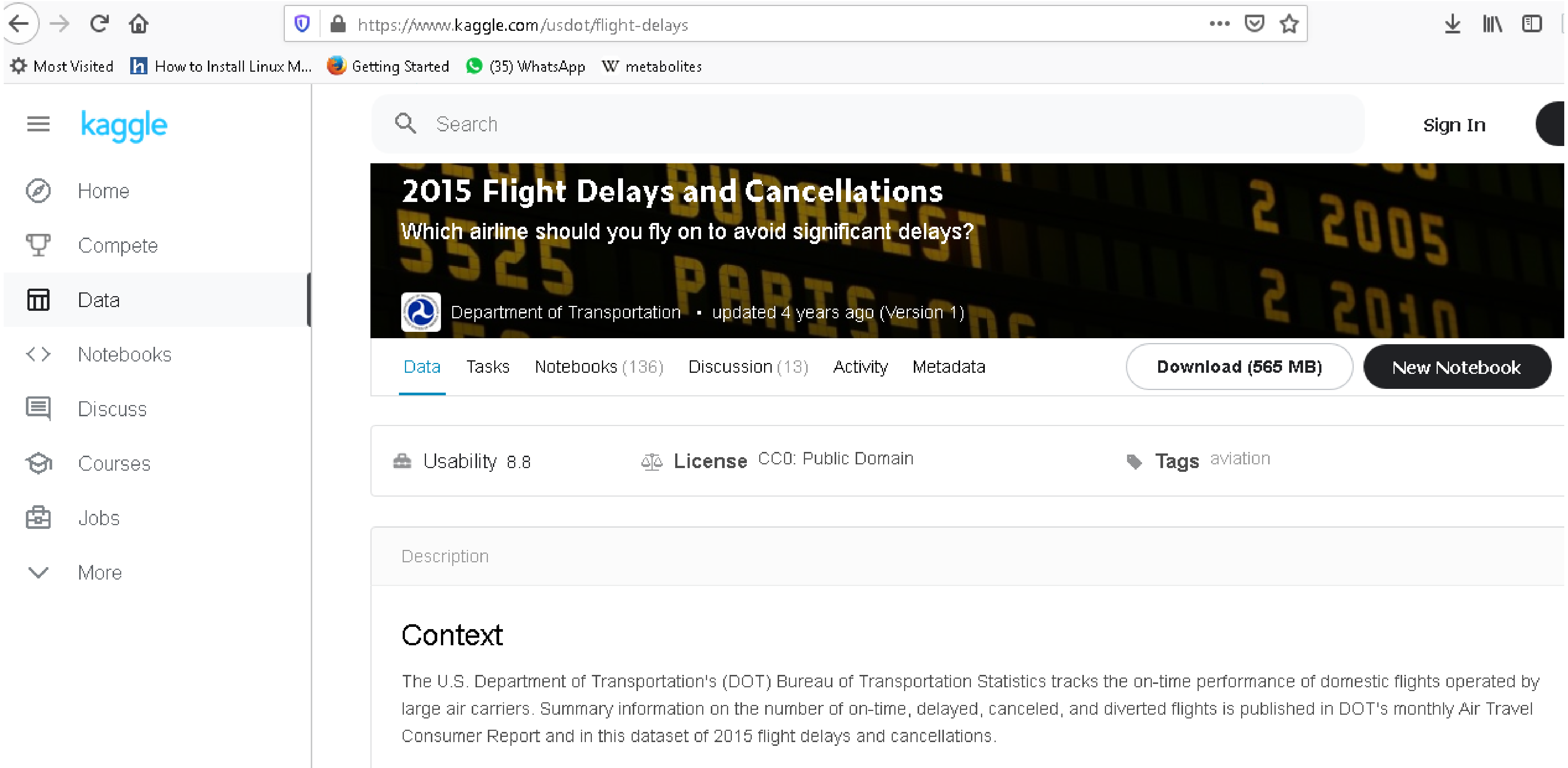
Identifying Frauds





```
harry   ->  pk :  2
ram   ->  pk :  4
shiv   ->  pk :  3
ford   ->  pk :  3
shayam   ->  pk :  2
dk   ->  pk :  1
john   ->  pk :  1
lee   ->  pk :  2
ck   ->  pk :  1
harry   ->  ck :  2
ram   ->  ck :  4
shiv   ->  ck :  3
ford   ->  ck :  3
shayam   ->  ck :  2
dk   ->  ck :  1
pk   ->  ck :  1
john   ->  ck :  1
lee   ->  ck :  2
```

Airlines & Airport Connectivity Network



2015 Flight Delays and Cancellations

Which airline should you fly on to avoid significant delays?

Department of Transportation • updated 4 years ago (Version 1)

Usability 8.8

License CC0: Public Domain

Tags aviation

Description

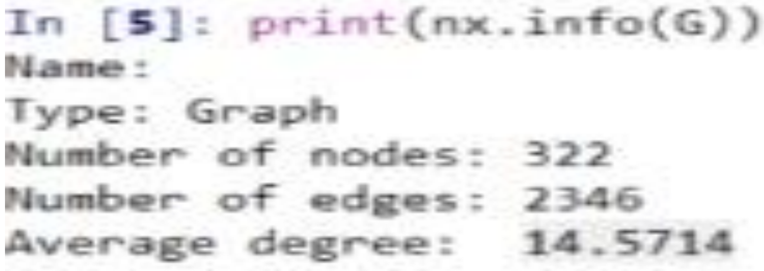
Context

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.

Airlines & Airport Connectivity Network

- **Busiest Airport**
- **Most Connectivity**
- **Shortest Route**
- **Least flights between Airports**
- **Identify Clusters**

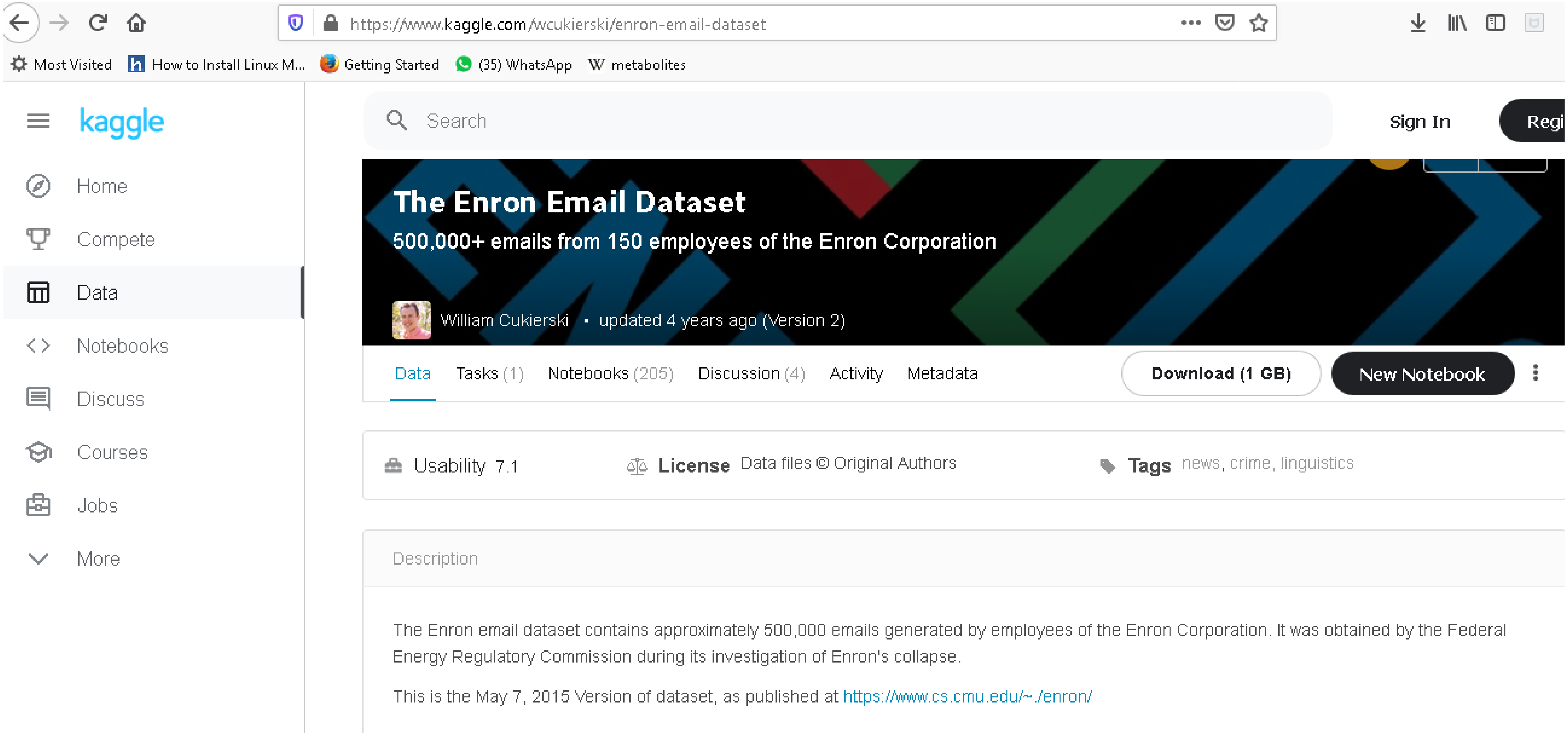
psg
wrg
qst
yak
bet
ktn
sit
ppg
cld
gum
brw
smx
ito
bli
spi
lnk
gfk
pub
ich
bro
mlb



```
In [5]: print(nx.info(G))
Name:
Type: Graph
Number of nodes: 322
Number of edges: 2346
Average degree:  14.5714
```

Is Transit from A → B → D → C ?

Enron Email Data



The Enron Email Dataset

500,000+ emails from 150 employees of the Enron Corporation

William Cukierski • updated 4 years ago (Version 2)

Usability 7.1

License Data files © Original Authors

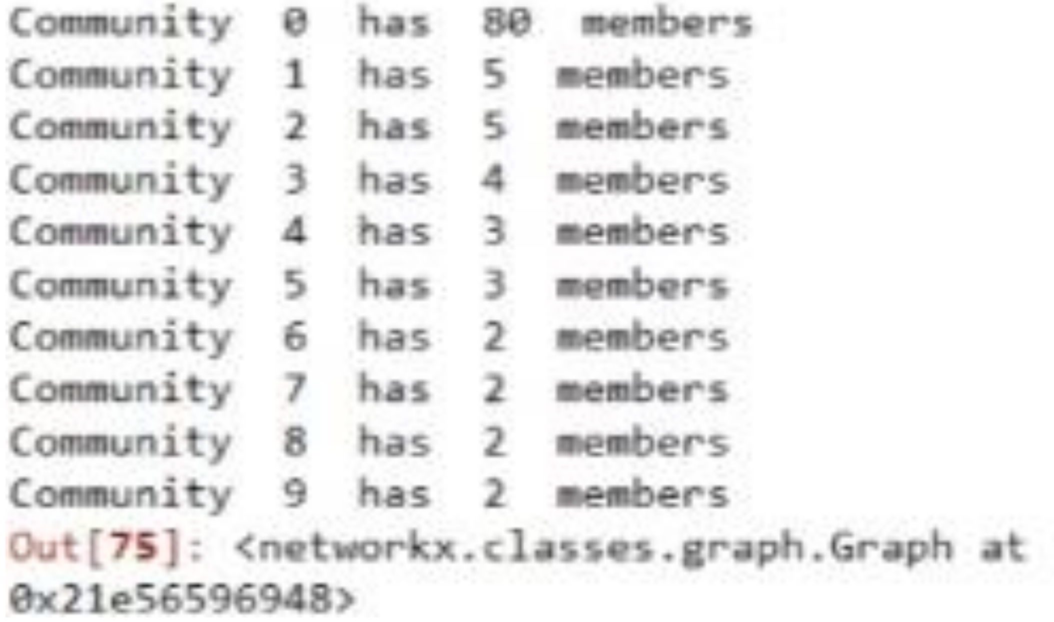
Tags news, crime, linguistics

Description

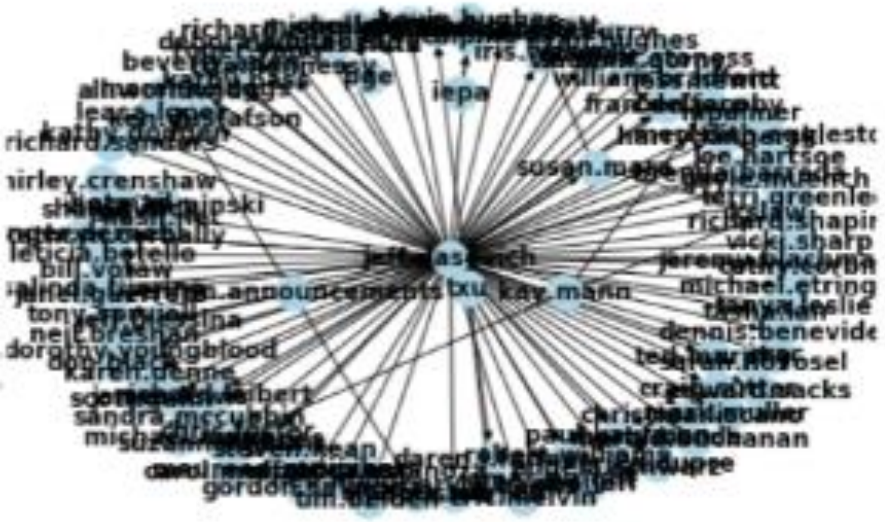
The Enron email dataset contains approximately 500,000 emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse.

This is the May 7, 2015 Version of dataset, as published at https://www.cs.cmu.edu/~./enron/

Enron Email Data
(Finding fraud people using Centrality Measures)



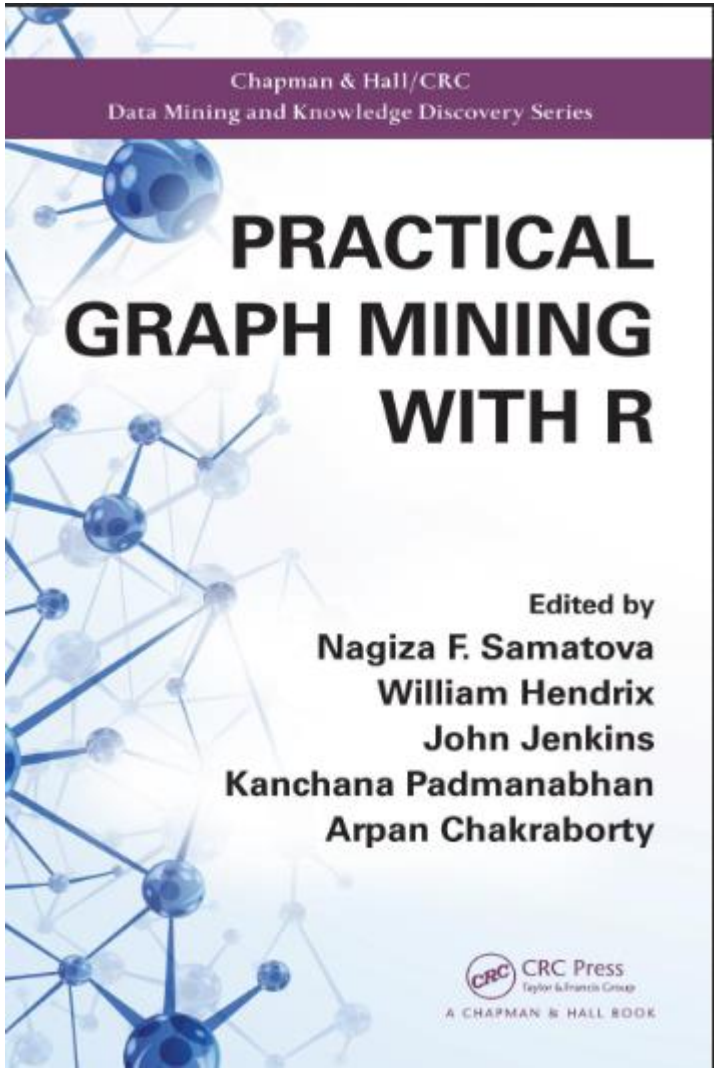
```
Community  0  has  80  members
Community  1  has  5  members
Community  2  has  5  members
Community  3  has  4  members
Community  4  has  3  members
Community  5  has  3  members
Community  6  has  2  members
Community  7  has  2  members
Community  8  has  2  members
Community  9  has  2  members
Out[75]: <networkx.classes.graph.Graph at
0x21e56596948>
```



To Summarize …

1. **Centrality analysis**: To identify the most central entities in your network, a very useful capability for influencer marketing.
2. **Path analysis**: To identify all the connections between a pair of entities, useful in understanding risks and exposure.
3. **Community detection**: To identify clusters or communities, which is of great importance to understanding issues in sociology and biology.
4. **Sub-graph isomorphism**: To search for a pattern of relationships, useful for validating hypotheses and searching for abnormal situations, such as hacker attacks.

Graph Mining Books

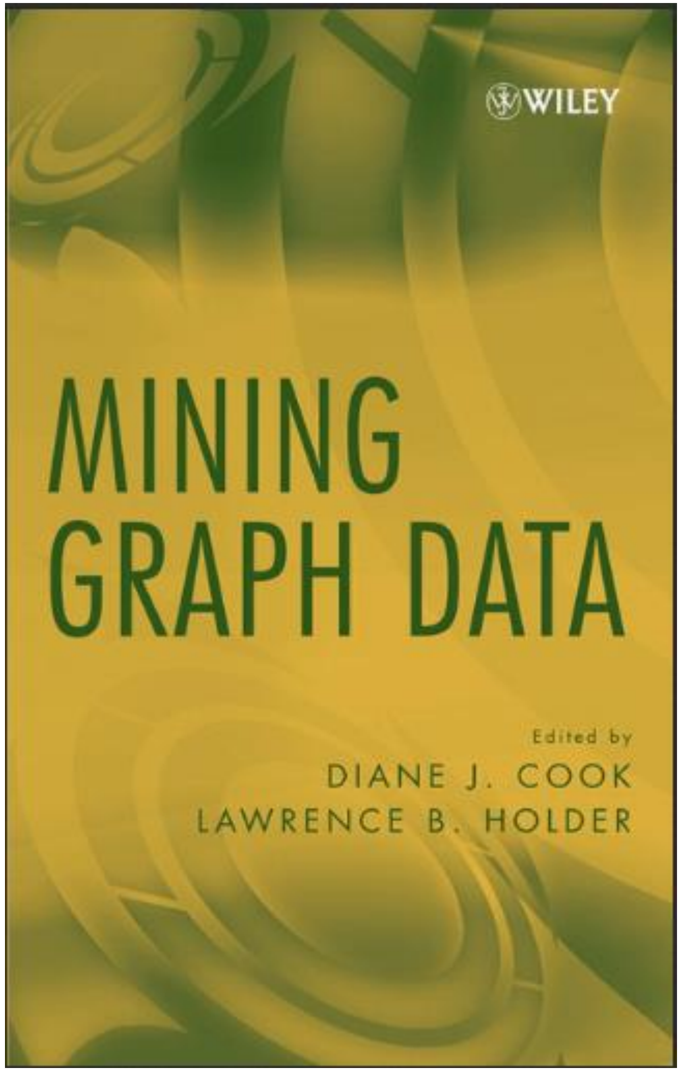


Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

PRACTICAL GRAPH MINING WITH R

Edited by
Nagiza F. Samatova
William Hendrix
John Jenkins
Kanchana Padmanabhan
Arpan Chakraborty

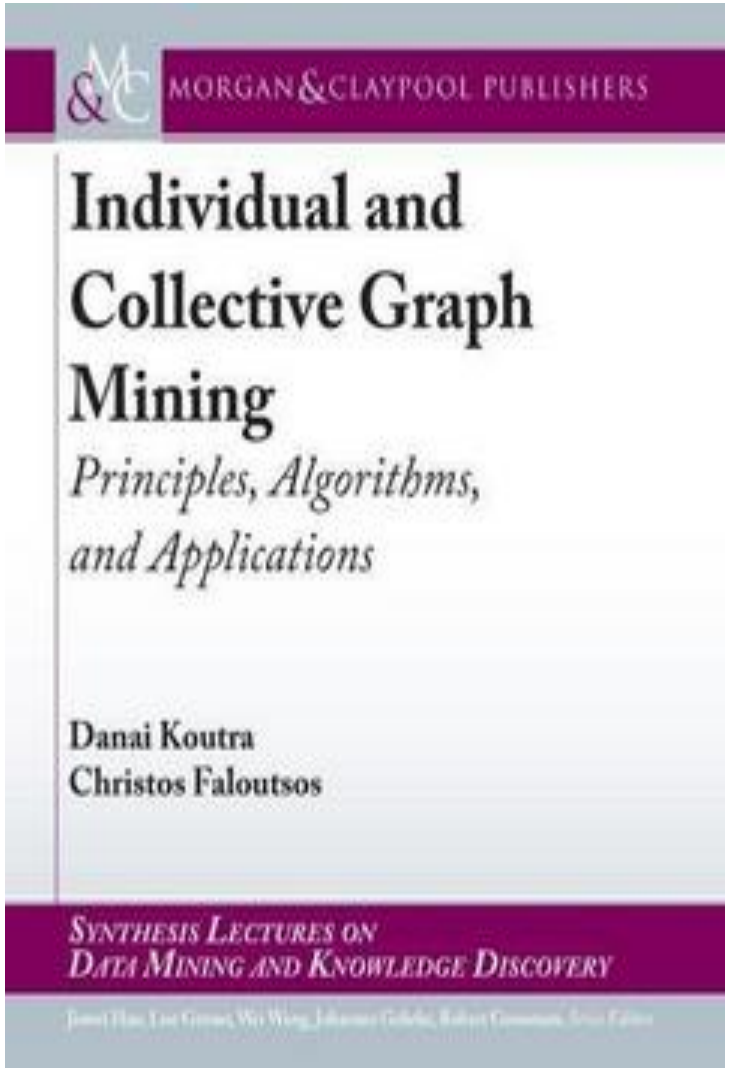
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



WILEY

MINING GRAPH DATA

Edited by
DIANE J. COOK
LAWRENCE B. HOLDER



M&C MORGAN & CLAYPOOL PUBLISHERS

Individual and Collective Graph Mining

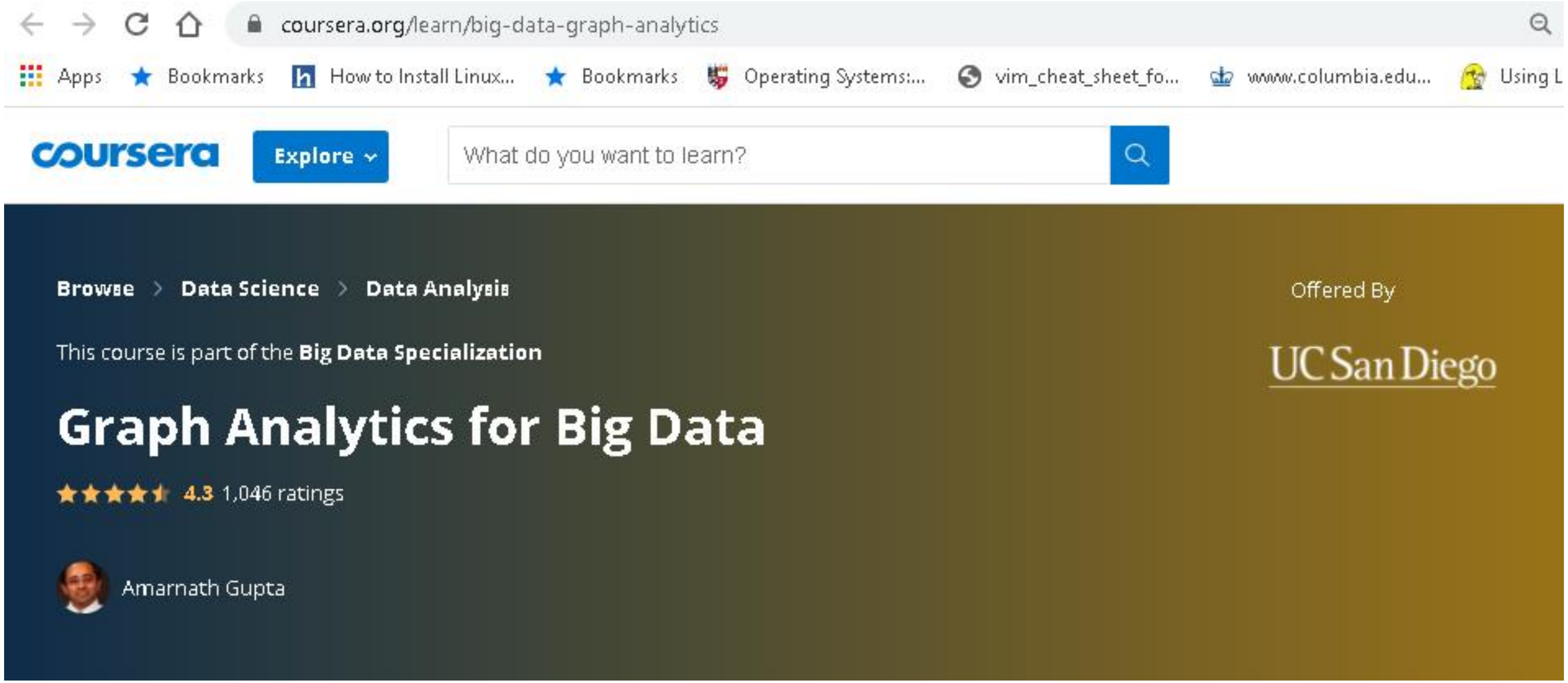
Principles, Algorithms, and Applications

Danai Koutra
Christos Faloutsos

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, Series Editors

Courses on Graph Analytics



Browse > Data Science > Data Analysis

This course is part of the **Big Data Specialization**

Graph Analytics for Big Data

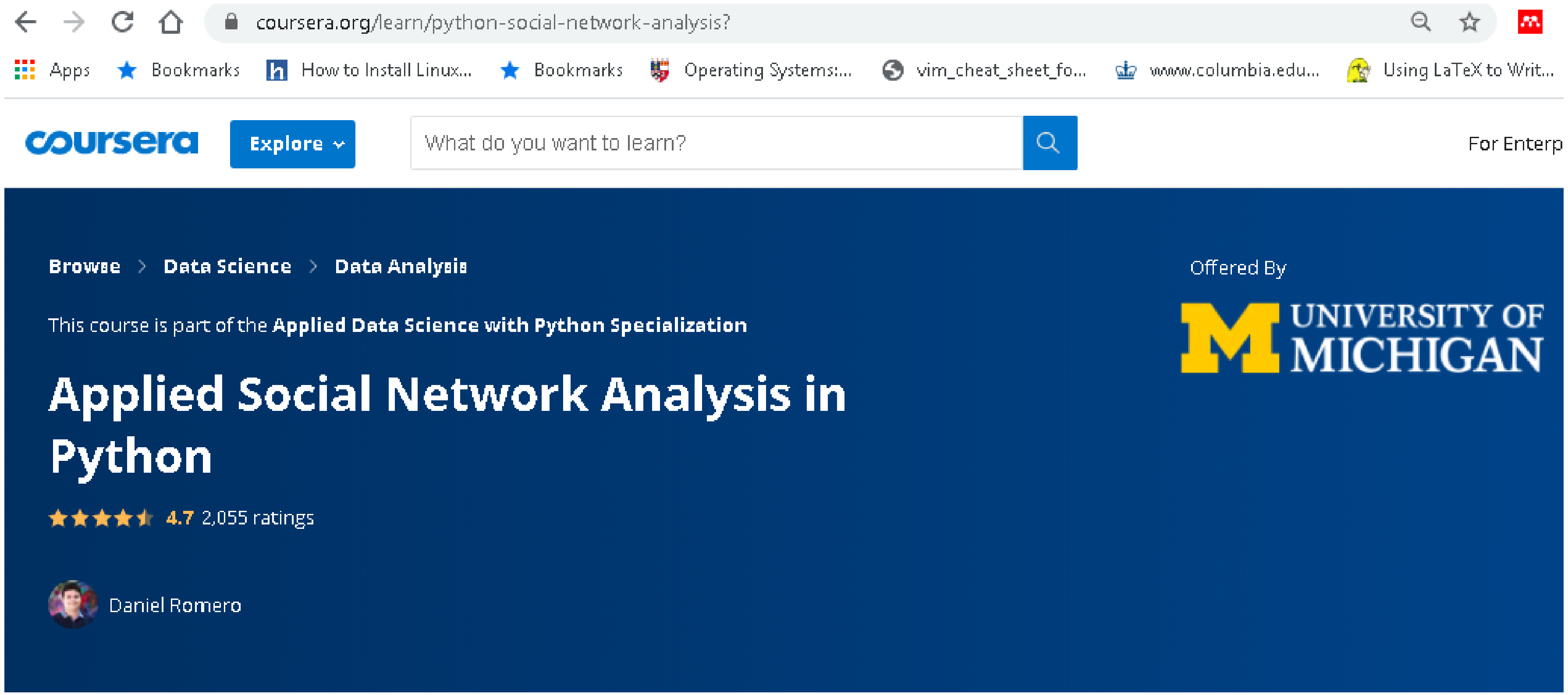
★★★★☆ 4.3 1,046 ratings

Amarnath Gupta

Offered By

UC San Diego

Courses on Graph Analytics



coursera.org/learn/python-social-network-analysis?

Browse > Data Science > Data Analysis

This course is part of the **Applied Data Science with Python Specialization**

Applied Social Network Analysis in Python

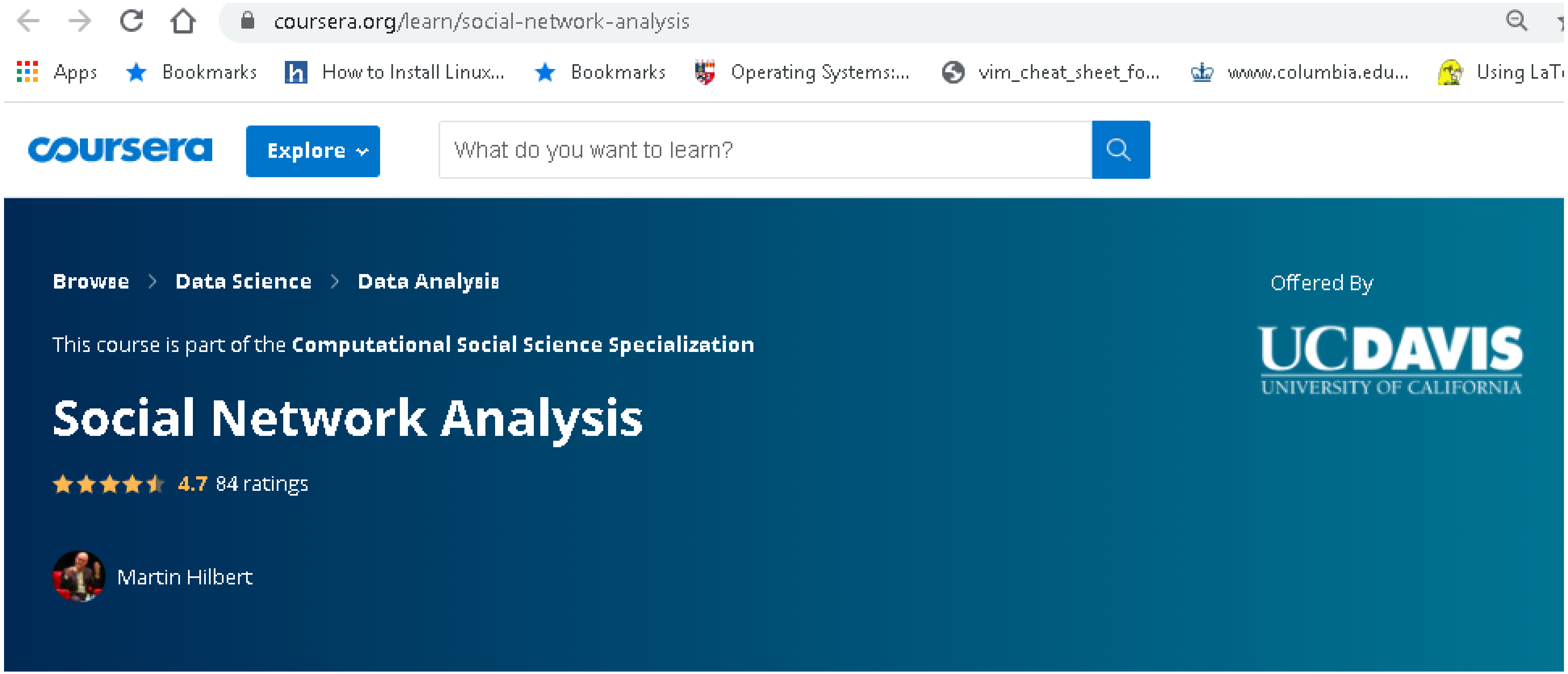
★★★★☆ 4.7 2,055 ratings

Daniel Romero

Offered By

UNIVERSITY OF MICHIGAN

Courses on Graph Analytics



coursera.org/learn/social-network-analysis

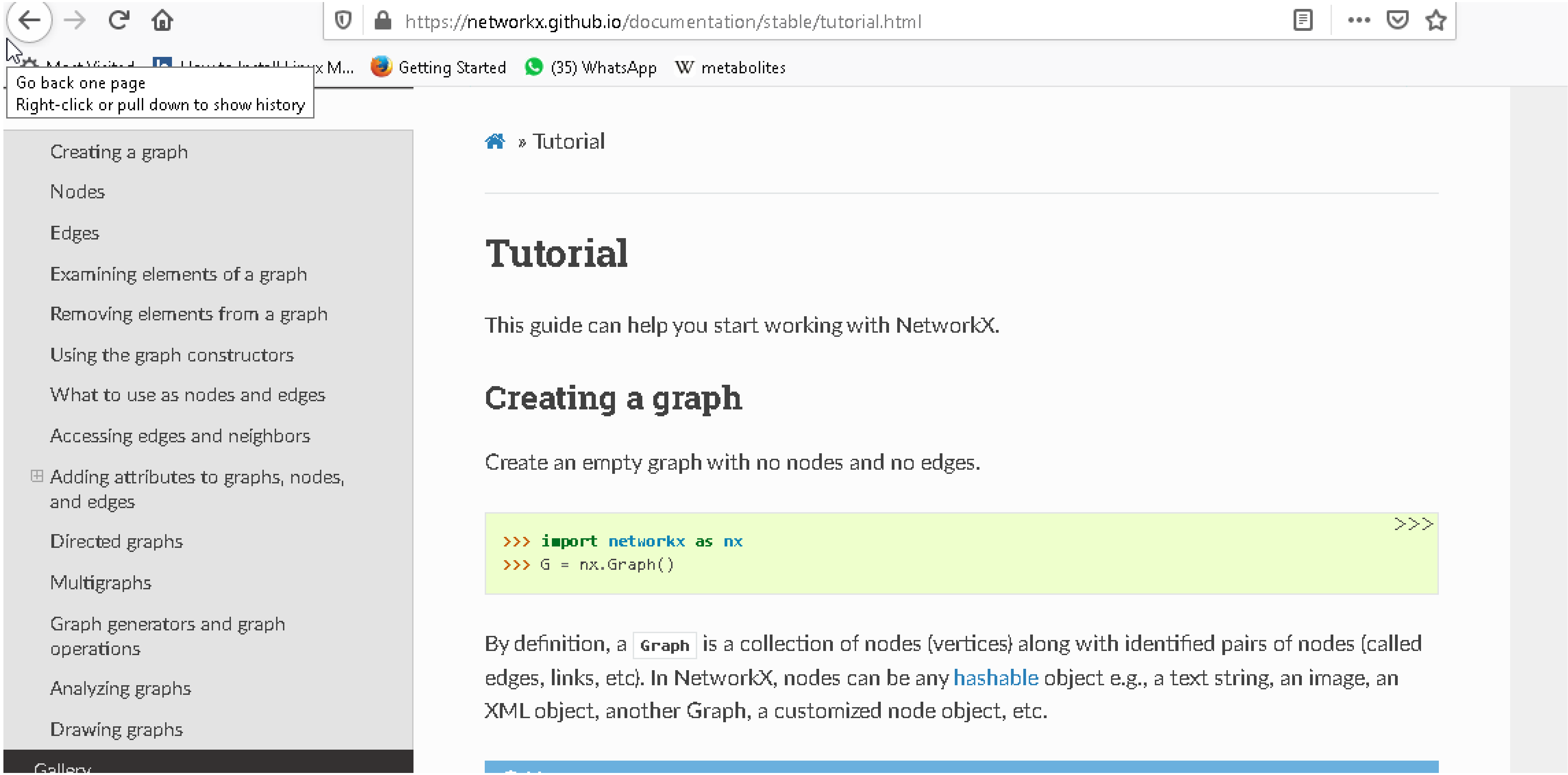
This course is part of the **Computational Social Science Specialization**

Social Network Analysis

★★★★½ **4.7** 84 ratings

Martin Hilbert

Tutorials on Networkx



https://networkx.github.io/documentation/stable/tutorial.html
Getting Started
(35) WhatsApp
metabolites
Go back one page
Right-click or pull down to show history

» Tutorial

Tutorial

This guide can help you start working with NetworkX.

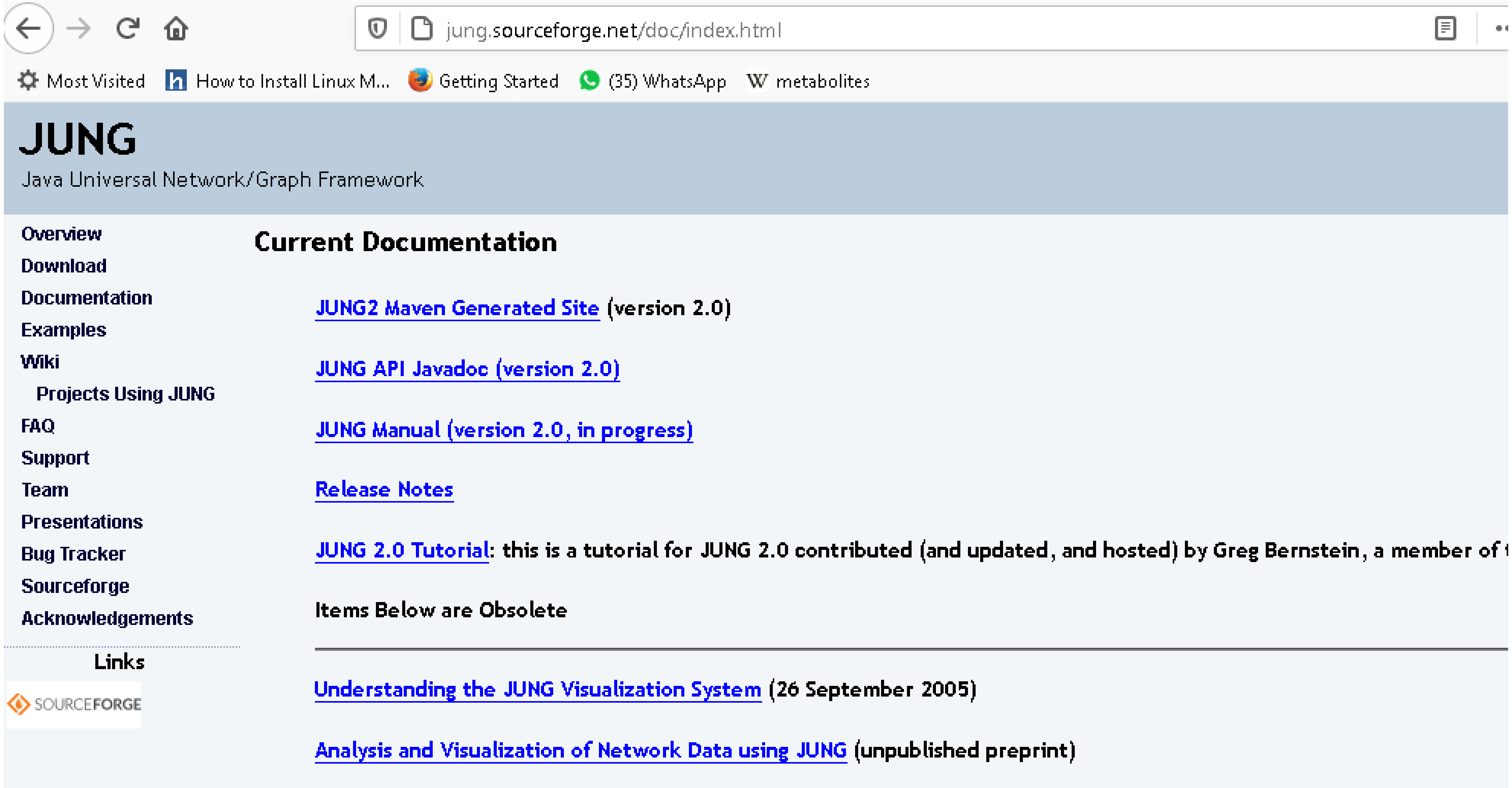
Creating a graph

Create an empty graph with no nodes and no edges.

```
>>> import networkx as nx
>>> G = nx.Graph()
```

By definition, a `Graph` is a collection of nodes (vertices) along with identified pairs of nodes (called edges, links, etc). In NetworkX, nodes can be any hashable object e.g., a text string, an image, an XML object, another Graph, a customized node object, etc.

Tutorials on JUNG



jung.sourceforge.net/doc/index.html
Most Visited
How to Install Linux M...
Getting Started
(35) WhatsApp
metabolites

JUNG

Java Universal Network/Graph Framework

Current Documentation

JUNG2 Maven Generated Site (version 2.0)

JUNG API Javadoc (version 2.0)

JUNG Manual (version 2.0, in progress)

Release Notes

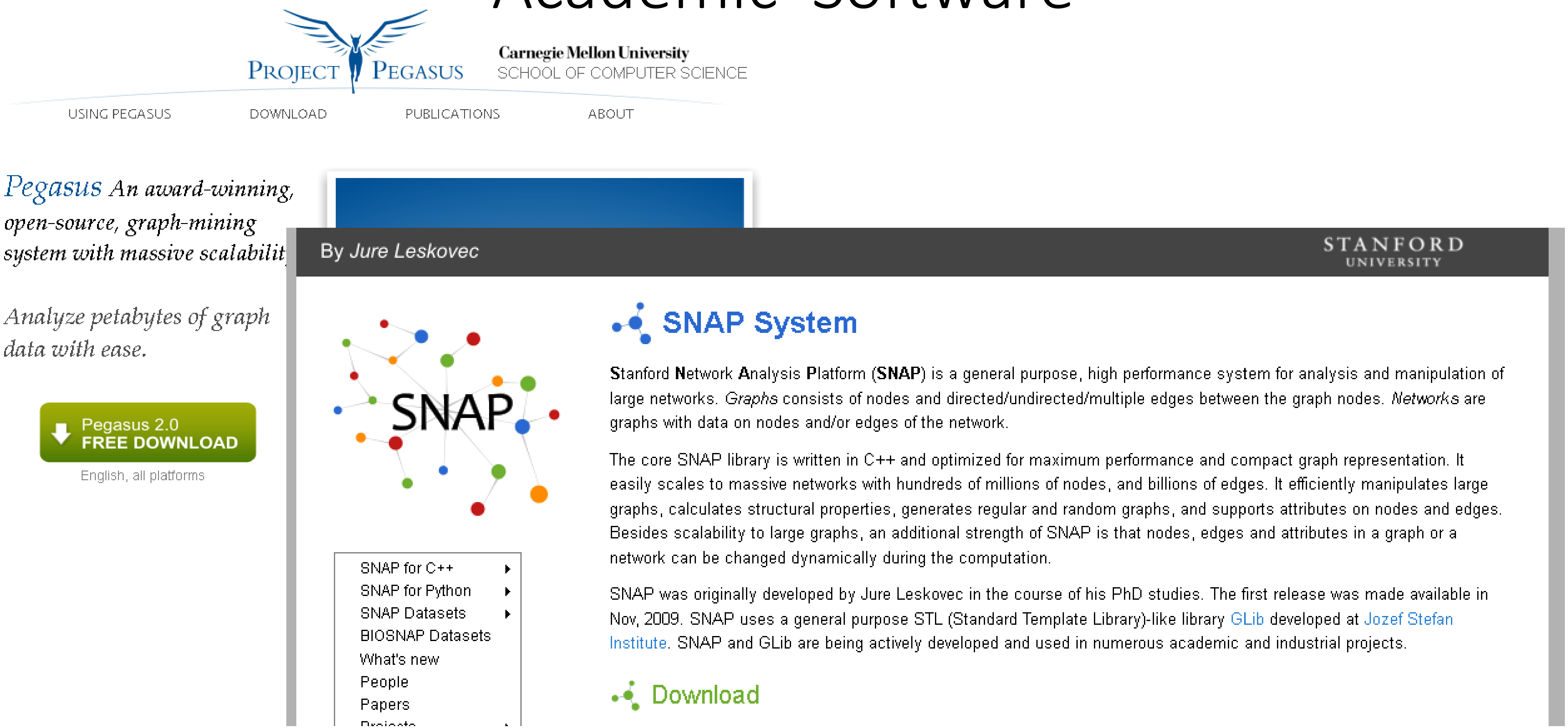
JUNG 2.0 Tutorial: this is a tutorial for JUNG 2.0 contributed (and updated, and hosted) by Greg Bernstein, a member of t

Items Below are Obsolete

Understanding the JUNG Visualization System (26 September 2005)

Analysis and Visualization of Network Data using JUNG (unpublished preprint)

Academic Software



Pegasus An award-winning, open-source, graph-mining system with massive scalabilit

Analyze petabytes of graph data with ease.

SNAP System

Stanford **N**etwork **A**nalysis **P**latform (**SNAP**) is a general purpose, high performance system for analysis and manipulation of large networks. *Graphs* consists of nodes and directed/undirected/multiple edges between the graph nodes. *Networks* are graphs with data on nodes and/or edges of the network.

The core SNAP library is written in C++ and optimized for maximum performance and compact graph representation. It easily scales to massive networks with hundreds of millions of nodes, and billions of edges. It efficiently manipulates large graphs, calculates structural properties, generates regular and random graphs, and supports attributes on nodes and edges. Besides scalability to large graphs, an additional strength of SNAP is that nodes, edges and attributes in a graph or a network can be changed dynamically during the computation.

SNAP was originally developed by Jure Leskovec in the course of his PhD studies. The first release was made available in Nov, 2009. SNAP uses a general purpose STL (Standard Template Library)-like library GLib developed at Jozef Stefan Institute. SNAP and GLib are being actively developed and used in numerous academic and industrial projects.

Download

Thank You