PERLOMBONGAN DATA GRAF

STQD6414 PERLOMBONGAN DATA

Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik

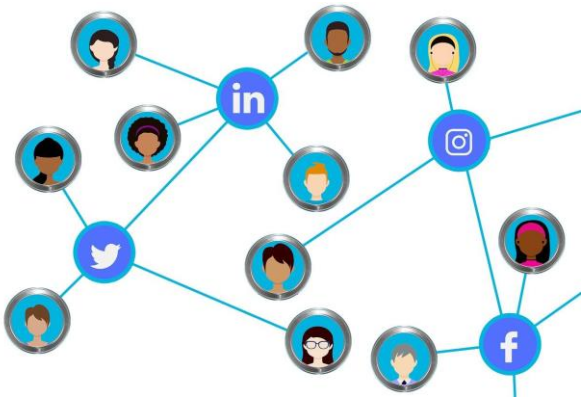
Universiti Kebangsaan Malaysia

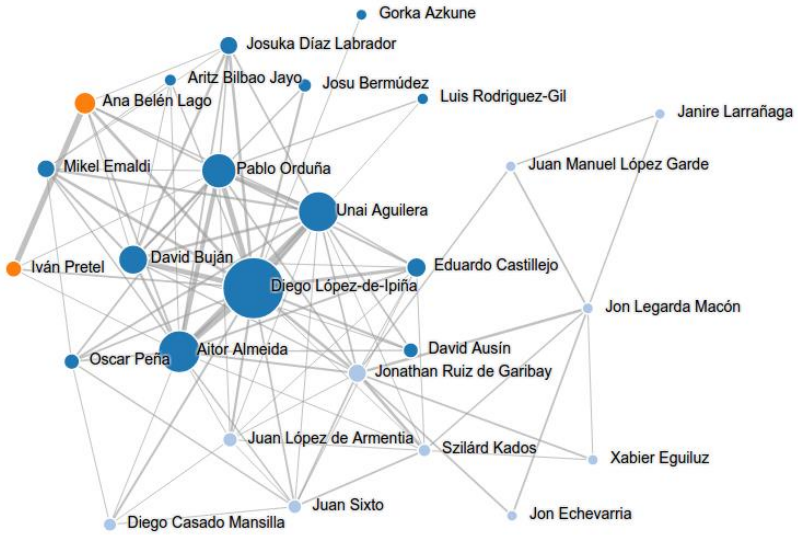
PENGENALAN:

- Graf ialah struktur data tak-linear yang terdiri daripada nod-nod/vertek (*nodes/vertices*) dan sisi-sisi (*edges*).
- Objektif perlombongan graf adalah untuk mengekstrak maklumat berguna daripada data yang diwakili dalam bentuk graf.
- Pada masa kini, data jenis graf terdapat di mana-mana, dan dipelbagai bidang.
- Contoh: Graf rangkaian sosial, graf Web, rangkaian keselamatan siber, rangkaian grid kuasa, pengurusan rantaian bekalan, rangkaian interaksi protein-protein, dan lain-lain.

PENGENALAN:

- Graf sering kali digunakan untuk menerangkan tentang pautan (*links*), perhubungan (*relationships*) atau kesalinghubungan (*interconnections*) antara suatu entiti.
- Contoh: dalam bidang sains sosial, nod dalam graf ialah individu dan pautan antara mereka adalah persahabatan atau hubungan kerjasama profesional, seperti yang boleh dicerap dari platform Facebook, LinkedIn, Instagram, Twitter, dll.
- Menerusi analisis data graf, kita boleh mendapatkan pemahaman yang lebih baik tentang ciri-ciri, tingkah laku atau trend interaksi antara entiti tertentu.



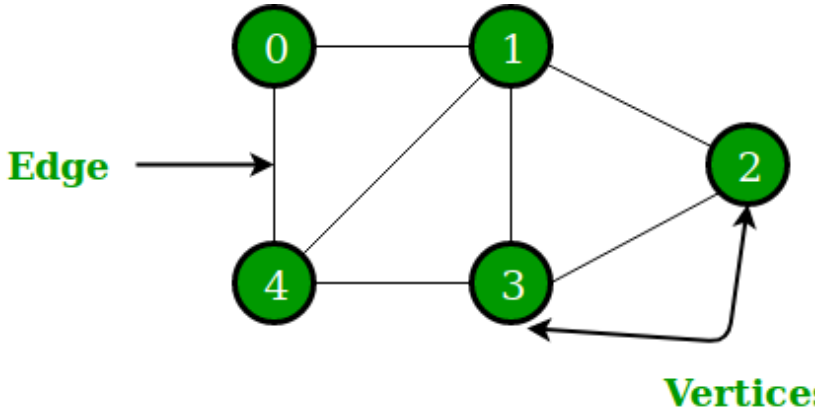


PENGENALAN:

- Untuk menganalisis struktur data graf, pengetahuan berkaitan teori graf adalah diperlukan.
- Teori graf adalah cabang ilmu matematik yang menerangkan berkaitan rangkaian titik-titik yang dihubungkan dengan garis-garis.
- Ia merupakan asas matematik yang digunakan untuk memodelkan hubungan berpasangan antara objek-objek.
- Umumnya, graf mewakili data berstruktur yang mengandungi verteks/nod (*vertices/nodes*) dan sisi (*edges*):

i) Nod graf mewakili maklumat berkaitan suatu entiti.

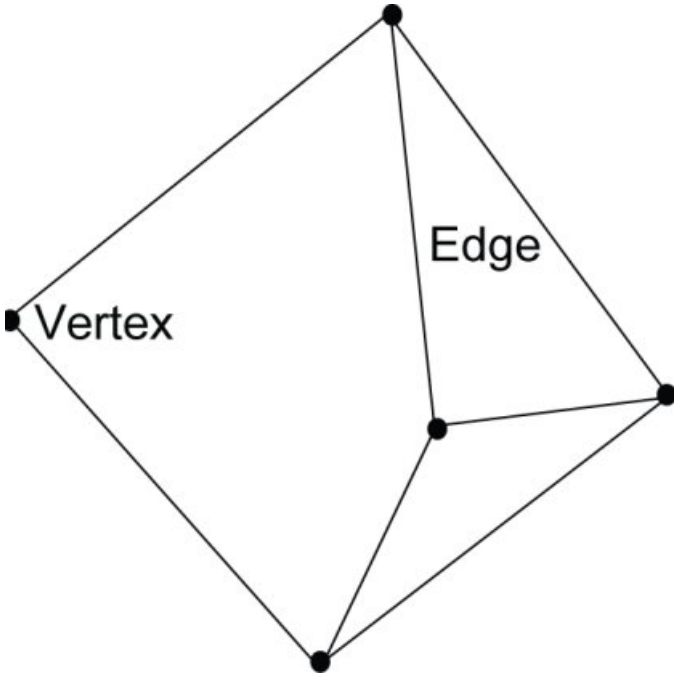
ii) Sisi graf mewakili hubungan antara maklumat antara entiti tersebut.



ASAS TEORI GRAF:

- Beberapa definisi penting:

i) Graf: Graf (G) terdiri daripada dua set: set verteks, ditunjukkan sebagai V(G), dan set sisi, ditunjukkan sebagai E(G).



ii) Sisi: Sisi dalam graf G ialah pasangan tak bertertib bagi dua verteks (v_1, v_2) sedemikian hingga $v_1 \in V(G)$ dan $v_2 \in V(G)$.

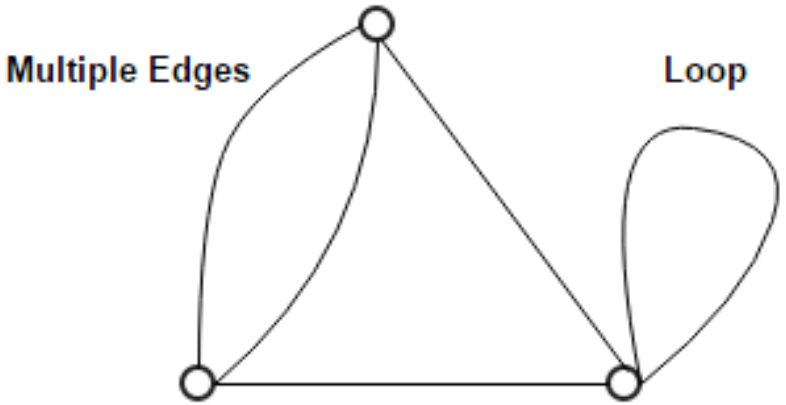
iii) Darjah (*Degree*): darjah(v), ialah bilangan kali verteks v berlaku sebagai titik akhir untuk sisi E(G).

ASAS TEORI GRAF:

iv) Gelung (*Loop*): Gelung ialah sisi yang menyambungkan suatu verteks kepada dirinya sendiri.

v) Sisi Berganda (*Multiple Edge*): Suatu sisi merupakan sisi berbilang jika terdapat sisi lain dalam E(G) yang bergabung dengan pasangan verteks yang sama.

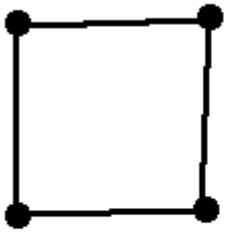
vi) Graf Mudah (*Simple Graph*): Graf tanpa gelung atau sisi berganda.



ASAS TEORI GRAF:

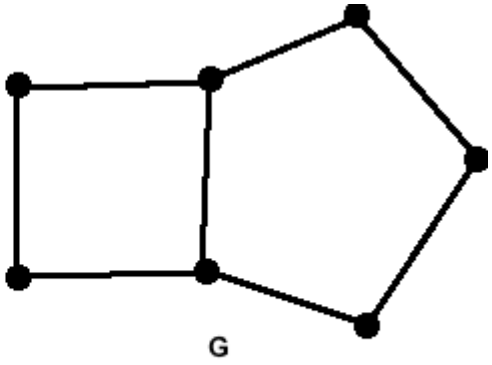
vii) Subgraf:

- Subgraf S bagi graf G ialah graf dengan set verteks V(S) yang merupakan subset bagi set verteks V(G), iaitu (V(S)⊆V(G)).
- Manakala, set sisi E(S) merupakan subset bagi set sisi E(G), iaitu (E(S)⊆E(G)).



S

(Subgraph of G)

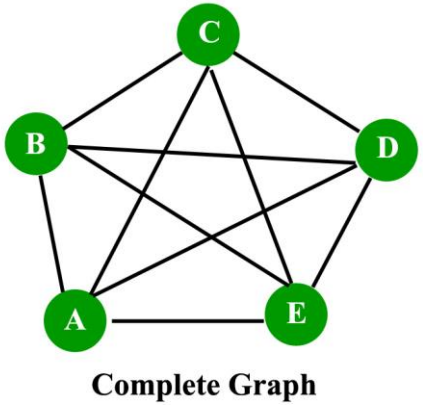


G

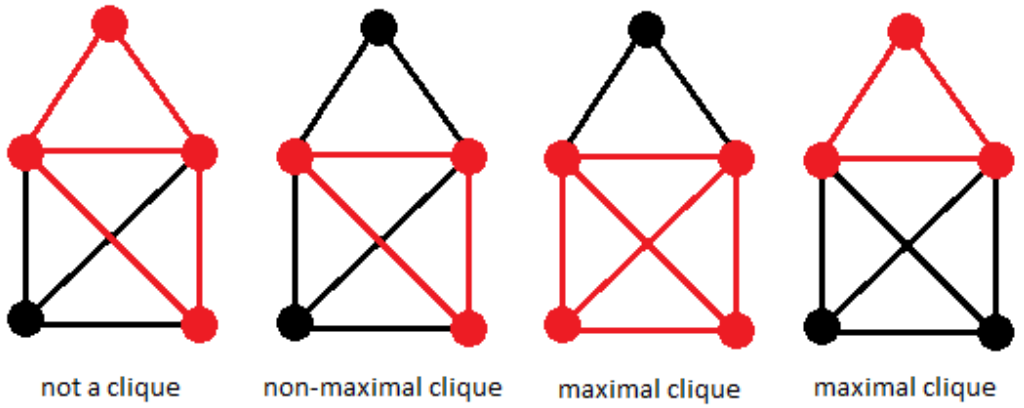
ASAS TEORI GRAF:

viii) Klik (*Clique*):

- Subset A⊆V dikatakan grap lengkap (*complete graph*) jika semua pasangan verteks dalam A adalah berhubung dengan suatu sisi.
- Iaitu, graf G=(V,E) adalah lengkap jika set verteks V adalah lengkap.
- Klik merupakan subset lengkap maksimum (*maximal complete subset*) jika subset lengkap tidak terkandung dalam subset lain yang lebih besar.
- Set bagi klik graf ditunjukkan sebagai C(G).



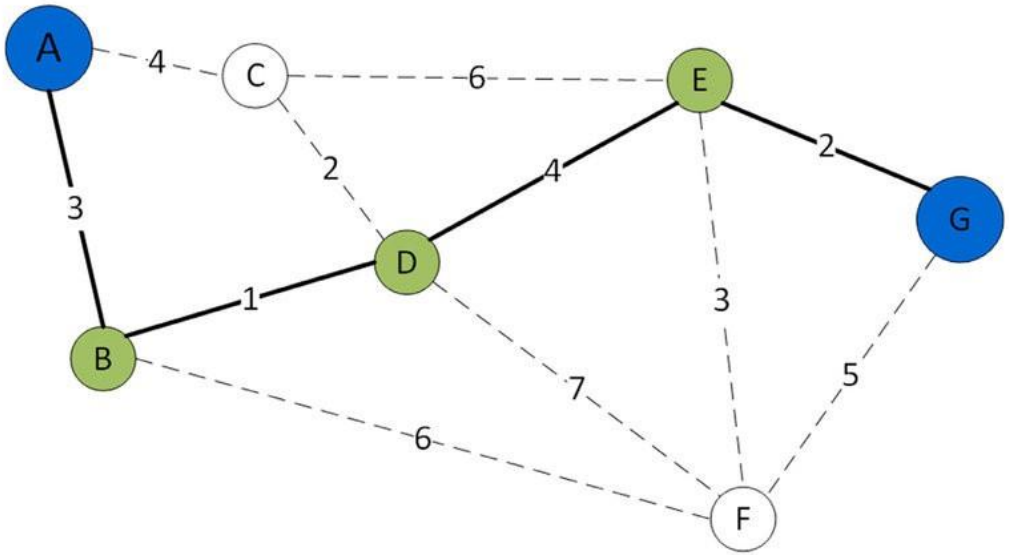
Complete Graph



ASAS TEORI GRAF:

ix) Laluan (*Path*) dan kitaran (*circle*):

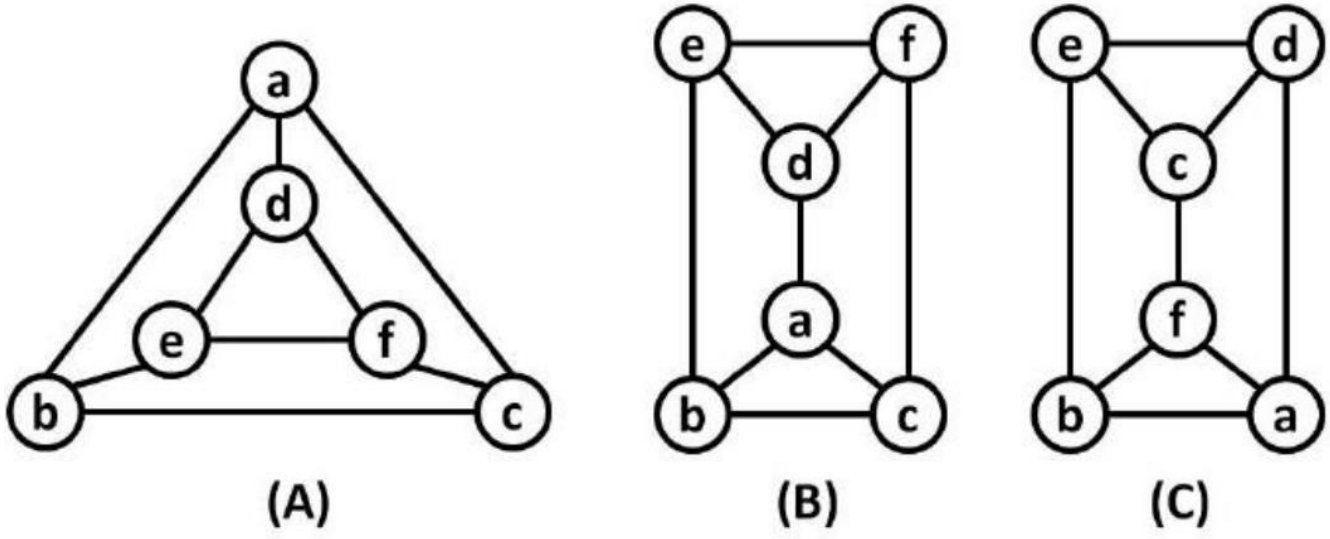
- Laluan (panjang *n*) antara verteks *α* dan *β* dalam graf tidak berarah ialah set verteks sedemikian hingga $\alpha=\alpha_0,\alpha_1,\ldots,\alpha_n=\beta$.
- Jika laluan $\alpha=\alpha_0,\alpha_1,\ldots,\alpha_n=\beta$ mempunyai sifat α=β, maka laluan tersebut disebut sebagai kitaran dengan panjang *n*.



ASAS TEORI GRAF:

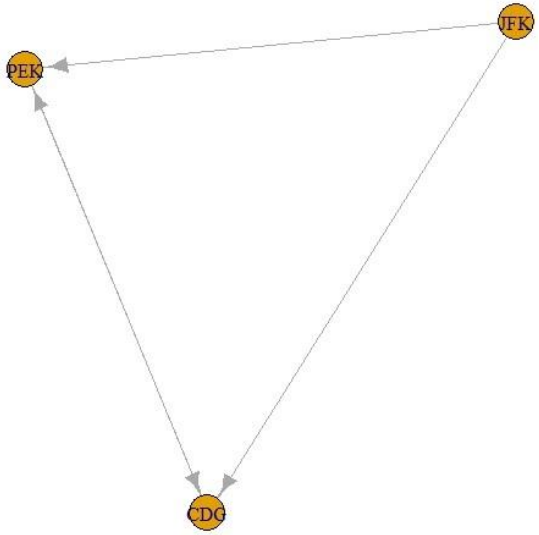
x) Graf Isomorfik (*Isomorphic*): Graf yang boleh wujud dalam bentuk yang berbeza tetapi mempunyai bilangan verteks, sisi dan juga ciri ketersambungan sisi yang sama

xi) Graf Automorfik (*Automorphic*): Graf yang mempunyai struktur yang sama, tetapi mempunyai tingkahlaku hubungan yang berbeza. Oleh itu, ianya bukanlah graf yang sama secara tepat.



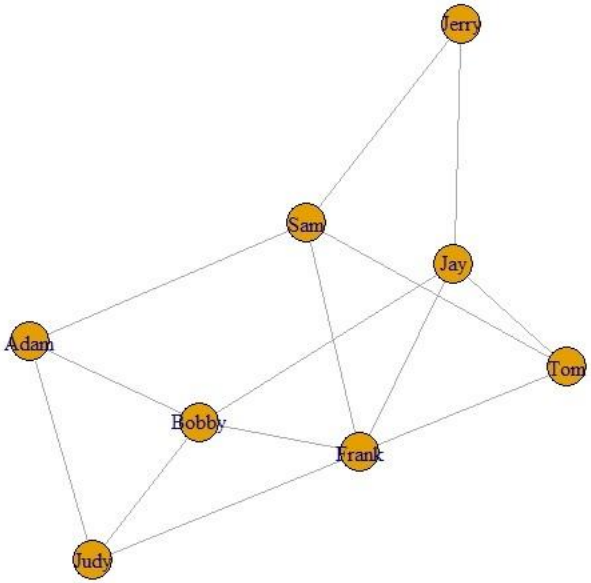
JENIS-JENIS GRAF:

- Terdapat pelbagai jenis graf antaranya ialah:



i) Graf Terarah (*Directed*) dan Tak-Terarah (*Undirected*):

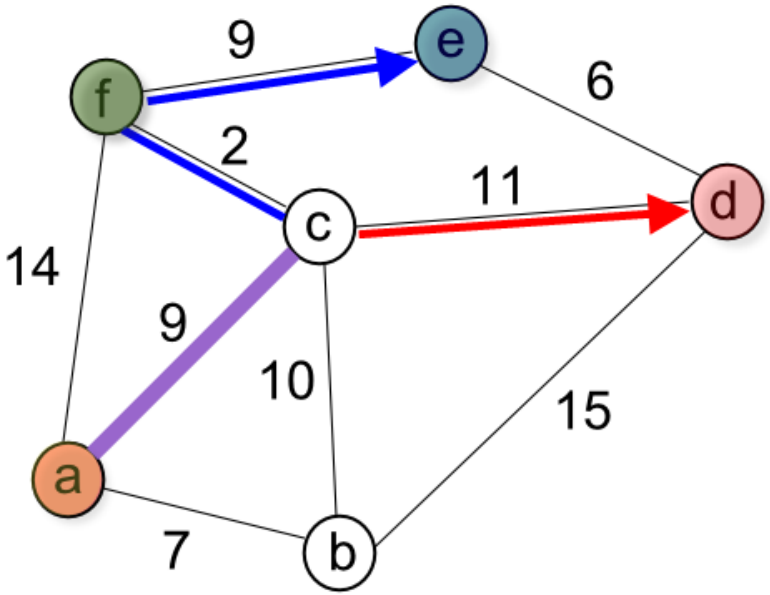
- Graf terarah mengandungi pasangan nod yang bertertib.
- Maka, graf terarah mempunyai sisi dengan arah yang khusus.
- Graf tak-terarah mengandungi pasangan nod yang tak-bertertib.
- Ini mengimplikasikan bahawa sisi graf tak-terarah tidak mempunyai arah tertentu.



JENIS-JENIS GRAF:

ii) Graf Berwajaran (*Weighted*) dan Tak-Berwajaran (*Unweighted*):

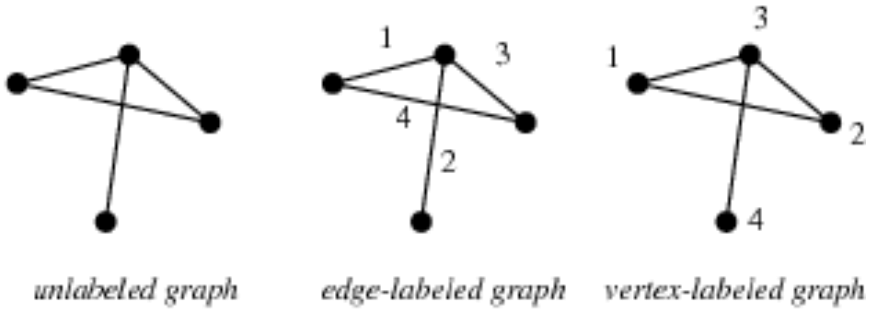
- Pemberat dalam graf mewakili magnitud hubungan antara nod-nod dan sisi-sisi.
- Graf yang mempunyai pemberat disebut sebagai graf berwajaran, dan begitu juga sebaliknya



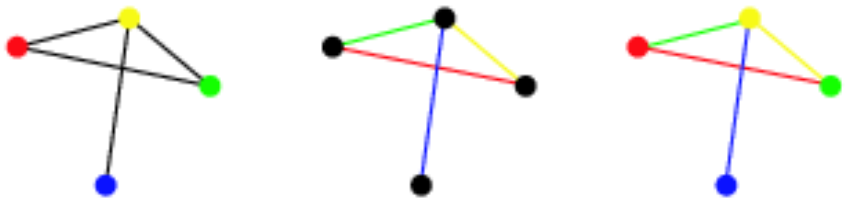
JENIS-JENIS GRAF:

iii) Graf Berlabel dan Tak-Berlabel:

- Graf tidak berlabel ialah graf dengan nod-nod atau sisi-sisinya tidak mempunyai sebarang petunjuk kecuali hanya melalui kesalinghubungannya.
- Manakala graf berlabel mempunyai beberapa petunjuk dalam nod atau sisinya.



unlabeled graph *edge-labeled graph* *vertex-labeled graph*



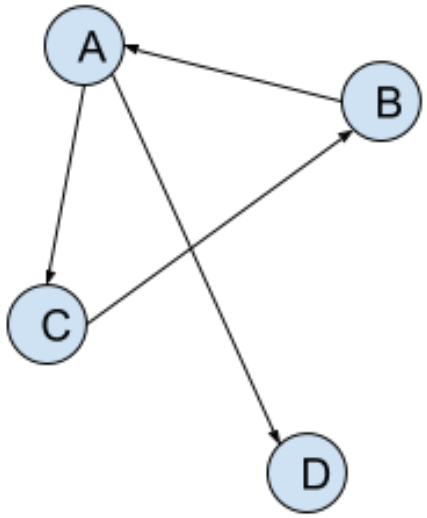
vertex-colored graph *edge-colored graph* *vertex- and edge-colored graph*

JENIS-JENIS GRAF:

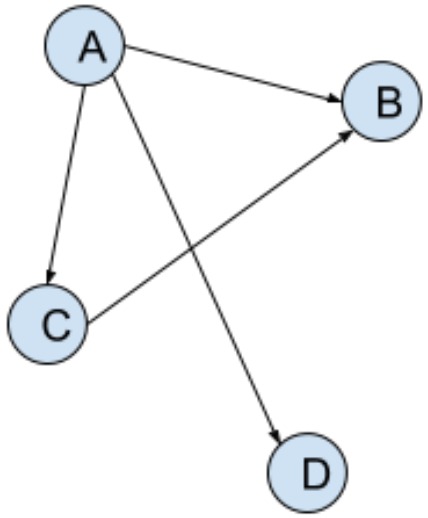
iv) Graf Berkitar (*Cyclic*) dan Tak-Berkitar (*Acyclic*):

- Graf yang mempunyai sekurang-kurangnya satu kitaran dipanggil graf berkitar.
- Graf tanpa kitaran disebut sebagai graf tak-berkitar (asiklik).

Cyclic Graph



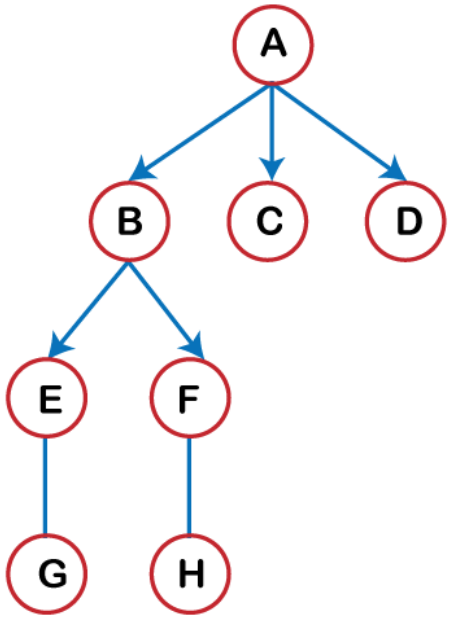
Acyclic Graph

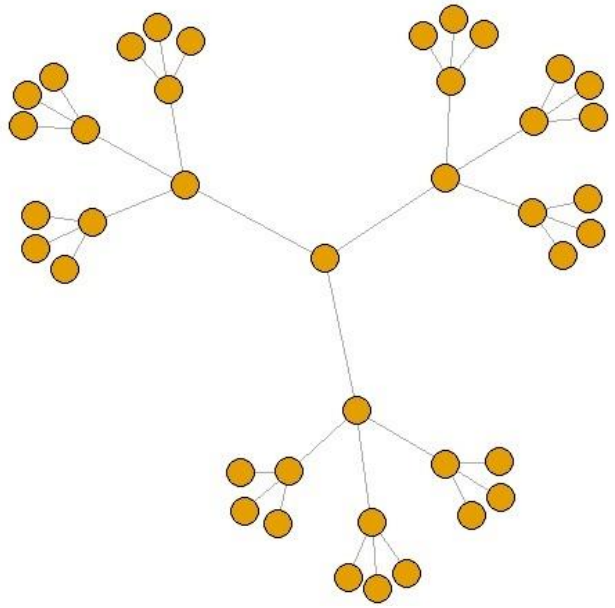


JENIS-JENIS GRAF:

v) Graf Pokok (*Trees*):

- Graf pokok merupakan graf tidak terarah dengan mana-mana dua verteks hanya bersambung dengan satu laluan (*path*) sahaja.
- Tiada kitaran berlaku dalam graf jenis ini.
- Ia juga dikenali sebagai graf tak-berarah asiklik yang berhubung.

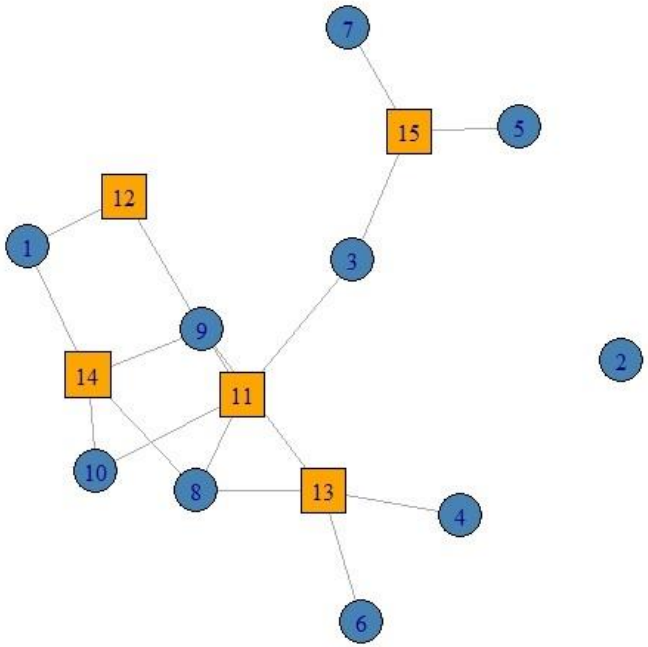




JENIS-JENIS GRAF:

vi) Graf Bipartit (*Bipartite*):

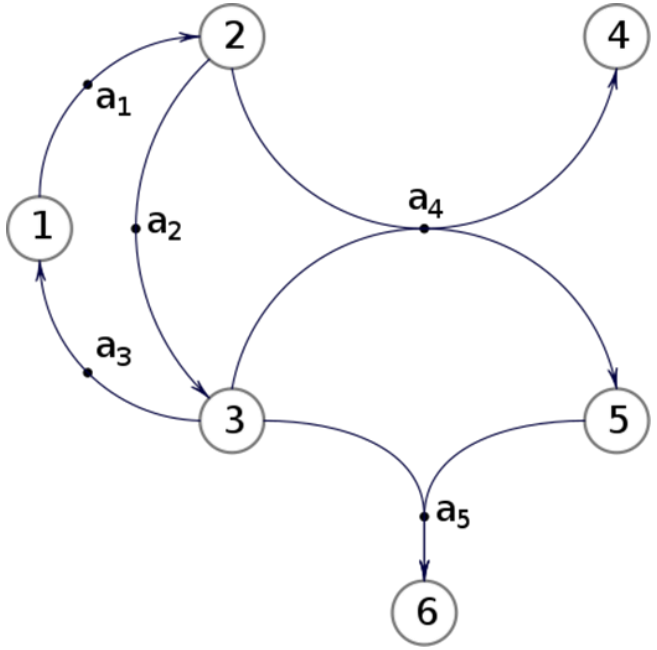
- Graf Bipartit ialah graf yang nod-nodnya boleh dibahagikan kepada dua set yang tak bersandar (U dan V)
- Setiap sisi (u,v) sama ada menghubungkan satu sisi dari U ke V atau satu sisi dari V ke U.
- Tiada sisi yang menghubungkan nod-nod dari set yang sama.
- Konsep graf bipartit boleh diitlakkan kepada graf multipartit.

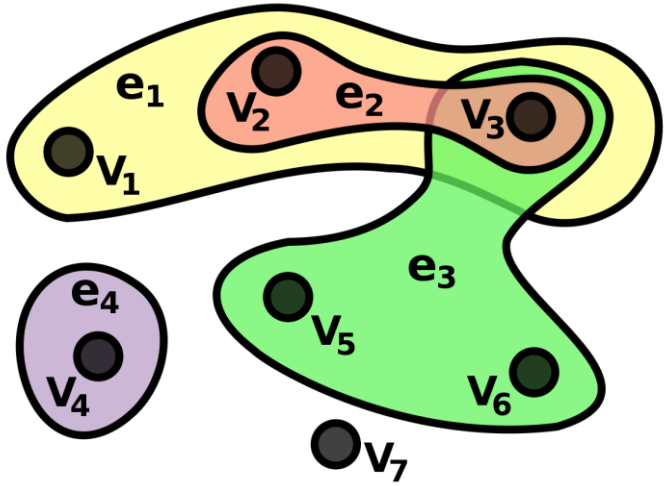


JENIS-JENIS GRAF:

vii) Hiper-graf (*Hypergraph*):

- Graf yang umum ialah 2-hipergraf (satu sisi menghubungkan 2 nod).
- Hiper-graf ialah graf teritlak yang mana sisinya boleh bergabung dengan sebarang bilangan nod-nod.
- Hiper-sisi (sisi teritlak) boleh berhubung dengan subset nod-nod berbanding graf bukan hiper yang hanya berhubung dengan 2 nod bagi satu sisi.
- K-hipergraf mempunyai semua hiper-sisi yang menyambungkan secara tepat k-nod.





PERWAKILAN DATA GRAF:

- Secara umumnya, data jenis graf disimpan dalam empat format asas:

i) Senarai Bersebelahan (*Adjacency lists*):

- Merupakan koleksi bagi senarai verteks tak bertertib.
- Setiap senarai tak bertertib menerangkan set jiran-jiran bagi suatu verteks dalam graf.

ii) Senarai Sisi (*Edge lists*):

- Merupakan jadual 2-kolum yang menyenaraikan semua pasangan vertex dalam graf.

```
$Adam
+ 3/8 vertices, named, from d339868:
[1] Judy  Bobby Sam

$Judy
+ 3/8 vertices, named, from d339868:
[1] Adam  Bobby Frank

$Bobby
+ 4/8 vertices, named, from d339868:
[1] Adam  Judy  Frank Jay

$Sam
+ 4/8 vertices, named, from d339868:
[1] Adam  Frank Tom   Jerry

$Frank
+ 5/8 vertices, named, from d339868:
[1] Judy  Bobby Sam   Jay   Tom
```

	V1	V2
1	Adam	Judy
2	Adam	Bobby
3	Adam	Sam
4	Judy	Bobby
5	Judy	Frank
6	Bobby	Frank
7	Bobby	Jay
8	Sam	Frank
9	Sam	Tom

PERWAKILAN DATA GRAF:

iii) Matriks Bersebelahan (*Adjacency matrix*):

- Matriks ini menunjukkan sama ada dua verteks dalam graf berhubung atau tidak.
- Jika terdapat pautan antara verteks "i dan j", maka indeks baris-lajur (i, j) akan ditandakan sebagai 1, jika tidak ianya ditanda . atau 0.

```
8 x 8 sparse Matrix of class "dgCMatrix"
       Adam Judy Bobby Sam Frank Jay Tom Jerry
Adam      .    1     1   1     .   .   .     .
Judy      1    .     1   .     1   .   .     .
Bobby     1    1     .   .     1   1   .     .
Sam       1    .     .   .     1   .   1     1
Frank     .    1     1   1     .   1   1     .
Jay       .    .     1   .     1   .   1     1
Tom       .    .     .   1     1   1   .     .
Jerry     .    .     .   1     .   1   .     .
```

MANIPULASI GRAF:

- Antara teknik penting manipulasi graf ialah:

i) Keluarkan verteks tertentu.

ii) Menjana subgraf.

iii) Menggabungkan graf-graf.

iv) Mengubah suai verteks data.

v) Mengubah suai sisi data.

ANALISIS JARINGAN DAN PAUTAN:

- Pautan (*link*) merujuk kepada hubungan antara dua entiti.
- Rangkaian merujuk kepada koleksi entiti dan pautan antara mereka.
- Perlombongan graf merupakan asas untuk analisis pautan dan rangkaian.

- Contoh:

i) Perlombongan graf boleh digunakan untuk mentafsir rangkaian melalui penentuan pengelompokan nod.

ii) Perlombongan graf berguna dalam menentukan ketumpatan nod-nod yang berhubung dalam data rangkaian.

iii) Perlombongan graf berguna dalam mengenalpasti struktur susun atur dalam data rangkaian.

ANALISIS PROMINENS NOD:

- Data rangkaian menarik untuk diselidiki kerana ianya mengandungi corak struktur yang khusus.
- Struktur ini akan mempengaruhi ciri-ciri nod/ahli dalam rangkaian.

- Contoh: Individu yang berhubung kepada ramai ahli rangkaian lain berkemungkinan melihat seluruh rangkaian dalam konteks berbeza berbanding individu yang terasing.
- Oleh itu, dengan mengesan lokasi ahli dalam rangkaian, kita boleh menilai prominens nod dalam data.
- Nod adalah prominens jika ikatannya (*ties*) lebih menonjol terhadap ahli lain dalam rangkaian.

ANALISIS PROMINENS NOD:

- Antara ukuran yang boleh digunakan untuk menjalankan analisis prominens nod:

i) Kepusatan Darjah (*Degree Centrality*):

- Berdasarkan ukuran ini, nod-nod yang mempunyai lebih banyak ikatan secara langsung (*direct ties*) adalah lebih prominens.

ii) Kepusatan Kedekatan (*Closeness Centrality*):

- Berdasarkan ukuran ini, nod lebih prominens jika ianya lebih dekat dengan semua nod-nod lain dalam rangkaian.

iii) Kepusatan Antara (*Betweenness Centrality*):

- Berdasarkan ukuran ini, nod lebih prominens jika lokasinya terletak 'antara' pasangan nod-nod lain dalam rangkaian.
- Laluan antara nod-nod lain perlu melalui nod yang prominens.

ANALISIS PROMINENS NOD:

iv) Skor vektor eigen kepusatan (*Eigenvector Centrality Scores*):

- Mengukur pengaruh transitif nod.
- Skor vektor eigen yang tinggi bermakna nod tersebut berhubung dengan nod-nod lain yang masing-masing mempunyai skor yang tinggi juga.

v) Skor kepusatan maklumat (*Information Centrality Scores*):

- Nod dengan pusat maklumat yang lebih tinggi mempunyai kawalan yang lebih kuat ke atas aliran maklumat dalam rangkaian.
- Ianya menunjukkan kewujudan sejumlah besar laluan pendek dalam struktur rangkaian.

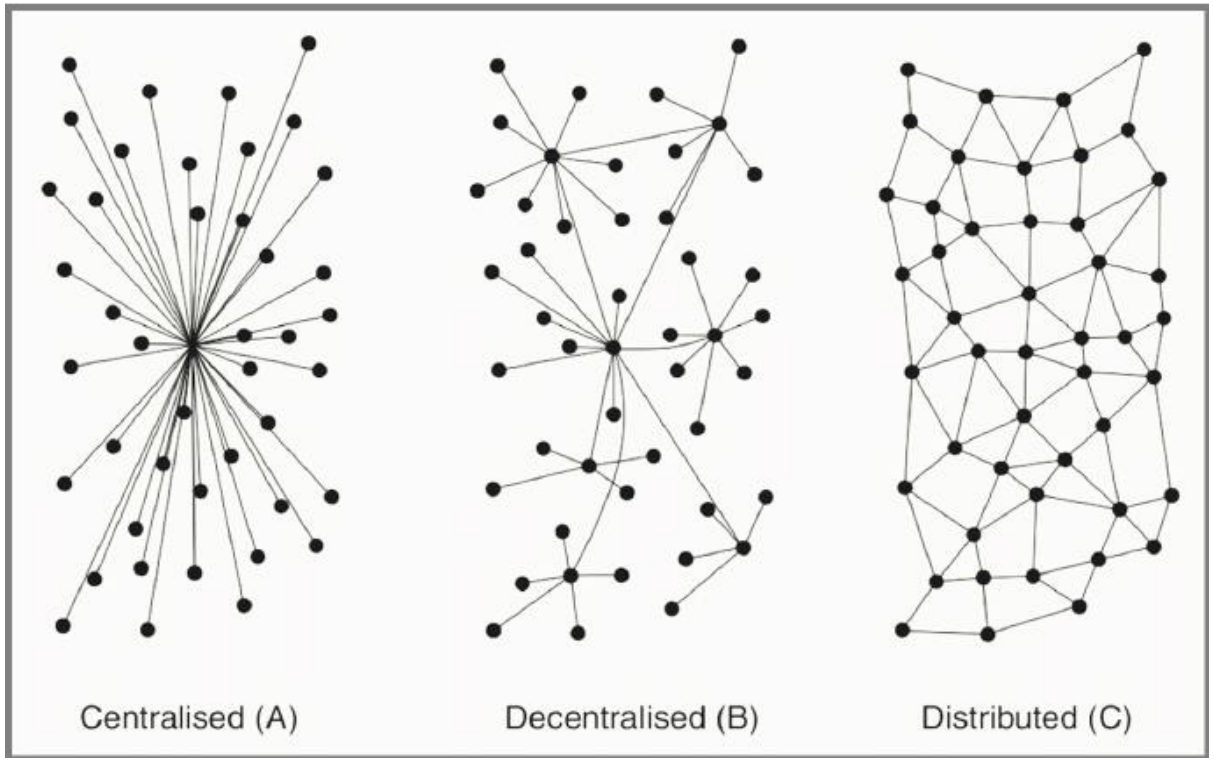
vi) Skor antara aliran (*Flow Betweenness Scores*):

- Mengukur jumlah aliran maksimum nod-nod tertentu.

ANALISIS PROMINENS NOD:

vi) Pemusatan (*Centralization*):

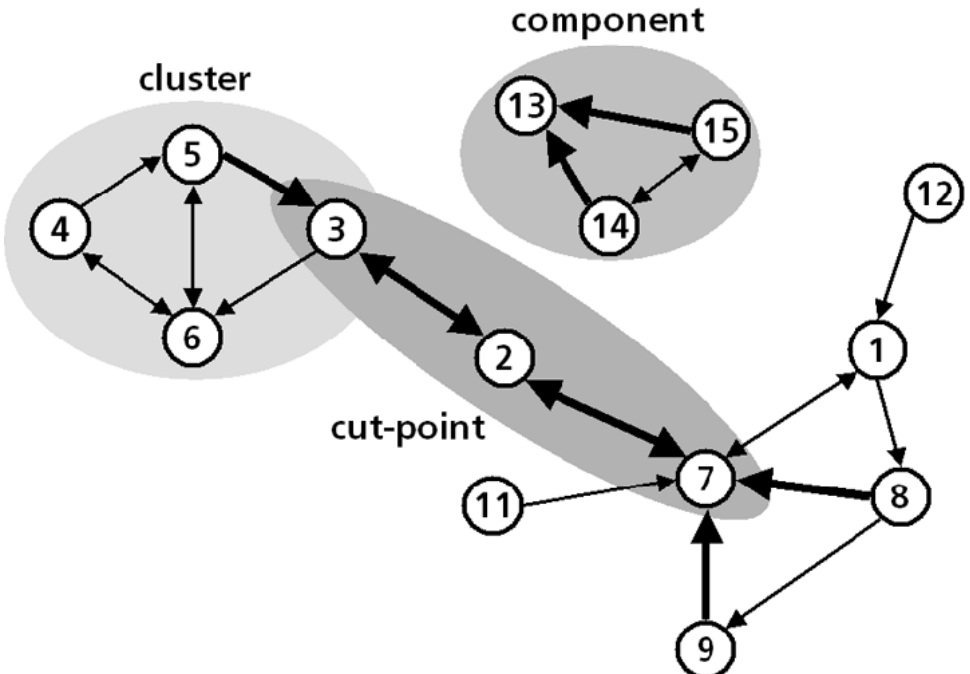
- Berdasarkan ukuran nod yang diberikan, kita boleh menganalisis tingkahlaku pemusatan rangkaian.
- Pemusatan merupakan ukuran variasi pusat rangkaian.



ANALISIS PROMINENS NOD:

vii) Titik potong (*Cutpoints*):

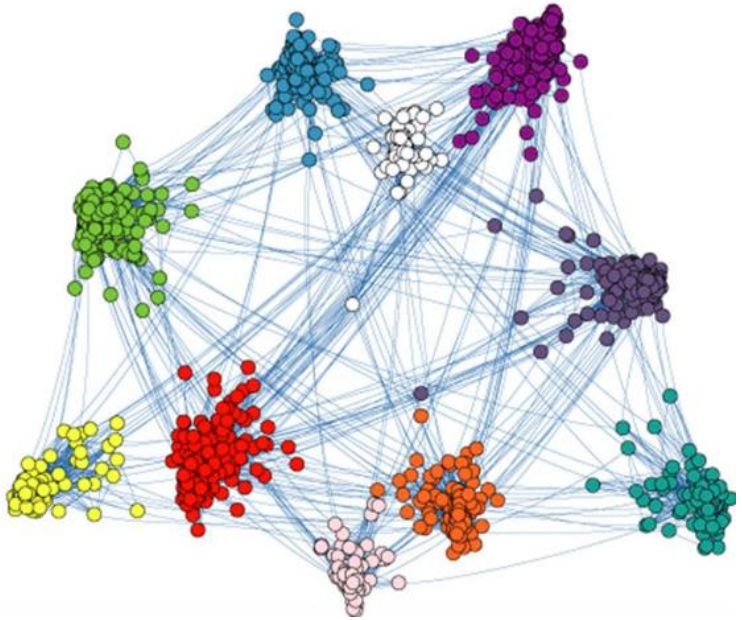
- Titik potong merujuk kepada nod yang jika kita keluarkannya, bilangan komponen dalam sub-rangkaian akan bertambah.
- Titik potong merupakan nod dengan kedudukan penting yang menghubungkan bahagian rangkaian yang berlainan.
- Jika nod titik potong dikeluarkan, itu akan mengakibatkan dua subset nod yang tidak akan dapat berkomunikasi antara satu sama lain.



ANALISIS SUB-KUMPULAN:

- Data rangkaian boleh dibentuk oleh beberapa sub-kumpulan padat yang berhubung hanya melalui ikatan yang agak lemah.

- Contoh: Sub-kumpulan bagi persahabatan boleh ditemui antara kenalan.
- Sub-kumpulan ini mengandungi maklumat yang berbeza antara satu sama lain.
- Oleh itu, untuk data rangkaian yang besar, adalah penting untuk dapat menentukan dan mengenal pasti sub-kumpulan tersebut untuk analisis lanjutan.



ANALISIS SUB-KUMPULAN:

- Dalam aplikasi sebenar, bagi rangkaian yang besar, kewujudan struktur sub-kumpulan umumnya sukar untuk dikesan dengan jelas.
- Oleh itu, analisis yang lebih sistematik perlu dilakukan untuk mengenalpasti kewujudan struktur sub-kumpulan.
- Struktur sub-kumpulan boleh dikesan berdasarkan konsep kejelekitan sosial (*social cohesion*).
- Jelekit sub-kumpulan (*Cohesive subgroups*) merujuk kepada set nod-not yang diikat bersama melalui ikatan yang kuat, kerap, dan dalam bentuk ikatan langsung.

- Dua jenis jelekit sub-kumpulan yang utama ialah:

i) Klik:

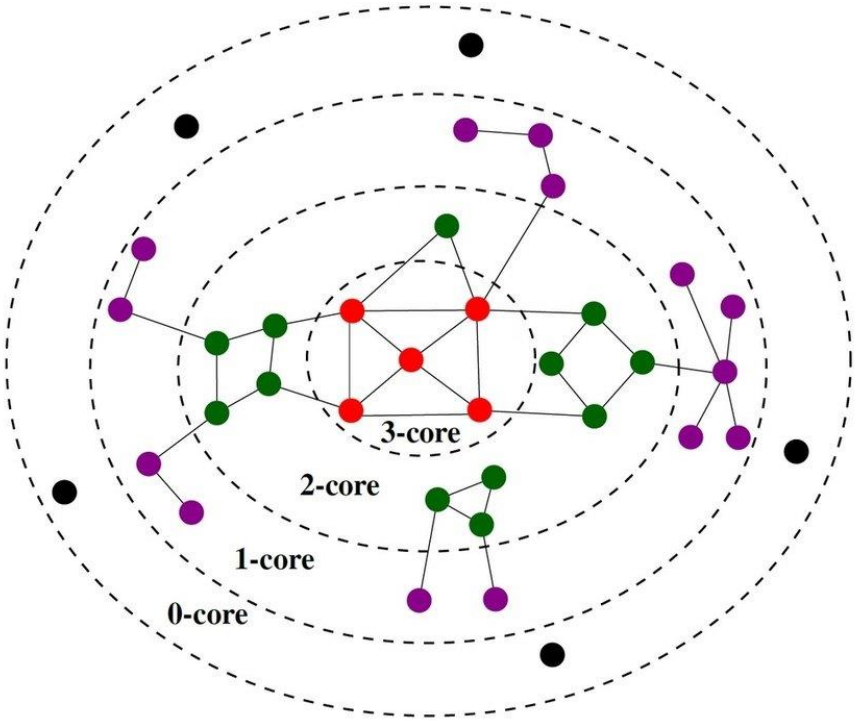
- Klik merupakan sub-kumpulan lengkap maksimum.
- Ia adalah subset nod-nod yang mempunyai semua kemungkinan ikatan antara mereka.

ANALISIS SUB-KUMPULAN:

ii) *k*-Teras (*k-Cores*):

- Klik kadangkala sukar untuk dikenalpasti kerana ia memerlukan syarat sub-graf yang lengkap maksimum.

- *k*-teras ialah ubahsuaian konsep klik yang merujuk kepada sub-graf maksimum dengan setiap verteks berhubung dengan sekurang-kurangnya *k* verteks lain dalam subgraf.

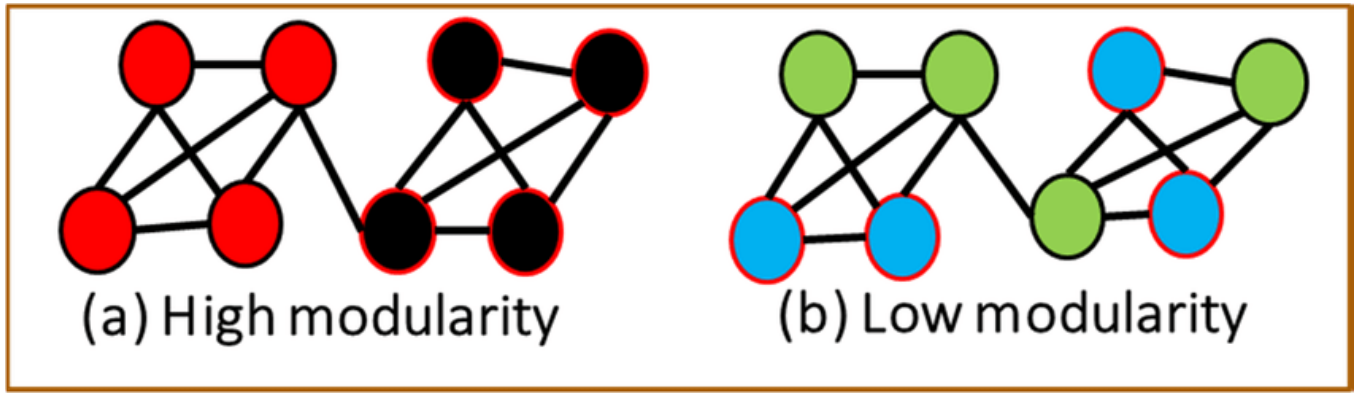


ANALISIS SUB-KUMPULAN:

Pendekatan lain untuk menanalisis struktur sub-kumpulan adalah berdasarkan teknik:

i) Kemodularan (*Modularity*):

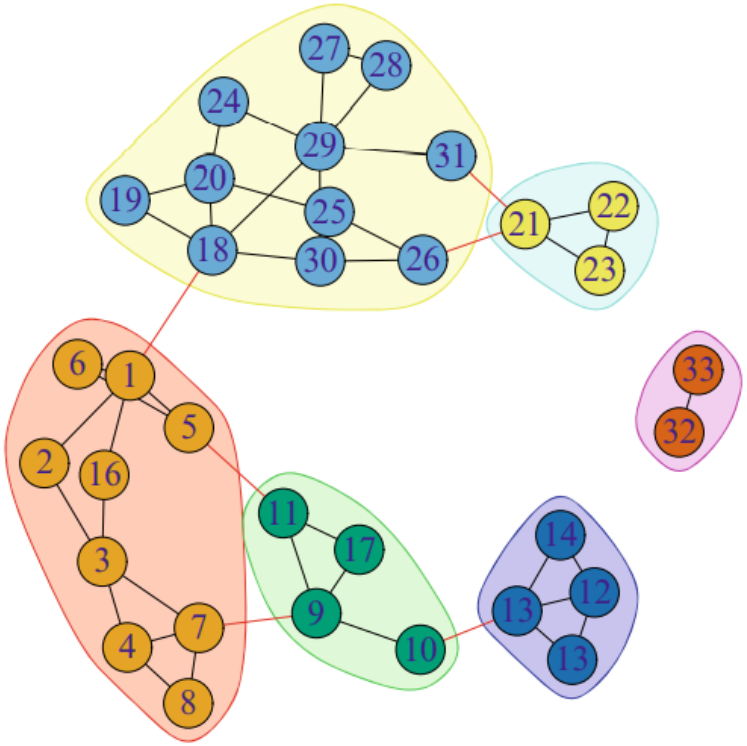
- Ini ialah ukuran struktur rangkaian, dengan nod-nod mempamerkan pengelompokan jika terdapat ketumpatan yang lebih besar dalam kelompok atau kurang ketumpatan di antara mereka.



ANALISIS SUB-KUMPULAN:

ii) Pengesanan komuniti (*Community Detection*):

- Komuniti dalam graf merujuk kepada subset nod-nod yang bersambung secara padat antara satu sama lain dan bersambung secara lemah dengan nod-nod dalam komuniti lain.



RUJUKAN:

- Brath, R., Jonker, D. (2015). *Graph analysis and visualization: Discovering business opportunity in linked data*. Wiley.
- Csardi, G., Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695.
- Gosnell, D., Broecheler, M. (2020). *The practitioner's guide to graph data: Applying graph thinking and graph technologies to solve complex problems*. O'Reilly Media
- Kolaczyk, E.D., Csárdi, G. (2020). *Statistical analysis of network data with R. Second Edition*. Cham: Springer.
- Luke, D.A. (2015). *A user's guide to network analysis in R*. Cham:Springer.
- Samatova, N.F., Hendrix, W., Jenkins, J., Padmanabhan, K., Chakraborty, A. (2014). *Practical graph mining with R*. Boca Raton: CRC Press.

TOPIK SETERUSNYA:

Perlombongan Data Web