PENGANTAR ANALISIS JEJARING

Budi Susanto (v.1.1)

Tujuan

- memahami metode centrality pada suatu graf untuk menemukan node yang paling berperan dalam jejaring.

Social Network

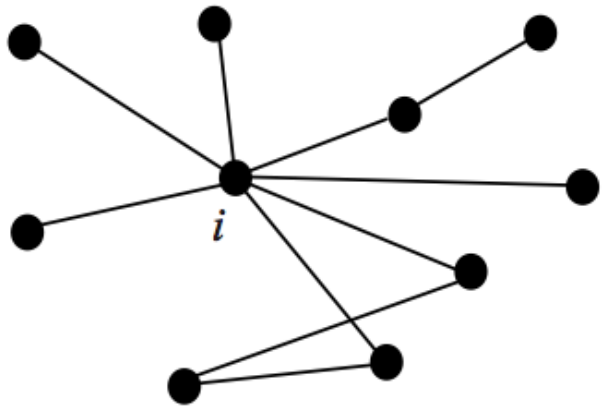
- Social network adalah studi terhadap entitas sosial (misalnya orang dalam suatu organisasi), dan interaksi serta relasi anar entitas tersebut.
- Interaksi dan hubungan dapat dinyatakan dengan suatu jaringan atau graf, di mana setiap vertex (node) menyatakan suatu hubungan.
- Dari jaringan tersebut, kita dapat mempelajari properti strukturnya, dan peran, posisi, dan martabat dari setiap aktor.
- Kita juga dapat menemukan berbagai macam bentuk sub-graf, seperti komunitas yang terbentuk dari sekelompok aktor.

Social Network untuk Web

- Social network analysis (SNA) bermanfaat juga untuk web karena web pada prinsipnya juga merupakan komunitas virtual
  - setiap halaman dapat diperlakukan sebagai aktor sosial dan setiap tautan sebagai sebuah hubungan antar aktor tersebut.
- Banyak hasil dari jejaring sosial dapat diadaptasi dan diperluas pemakaiannya dalam kontek Web.

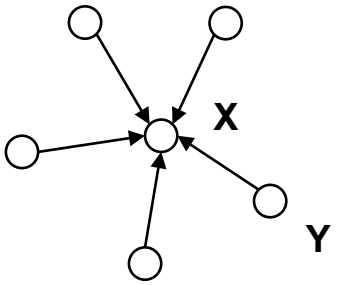
Centrality

- Dalam kontek suatu organisasi, seseorang dengan hubungan atau komunikasi yang ekstensif dengan banyak orang lain dalam organisasi dinilai lebih penting daripada orang lain yang memiliki kontak lebih sedikit
- Tautan atau hubungan dapat juga disebut sebagai ikatan (*ties*).
- Seorang aktor pusat terlibat dalam banyak ikatan.

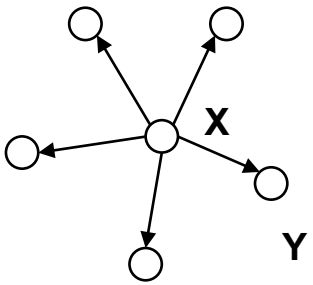


Centrality

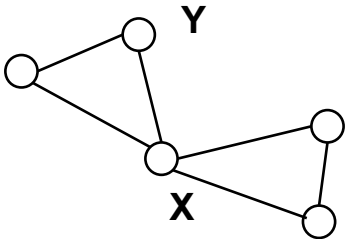
- Degree centrality
  - out-links
  - in-links
- Closeness centrality
- Betweeness centrality



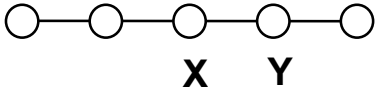
indegree



outdegree



betweenness



closeness

Degree Centrality

- dimisalkan total jumlah aktor dalam suatu jaringan adalah ***n***.
- Dalam undirected graph:
  - degree centrality dari seorang aktor i (dinyatakan sebagai $C_D(i)$) adalah derajat (jumlah edge) dari node aktor, dinyatakan sebagai *d(i),* dinormalisasikan dengan nilai maksimum degree, *n-1*.
  - Nilai dari pengukuran tersebut adalah 0 – 1, di mana *n-1* adalah nilai maksimum dari *d(i)*.

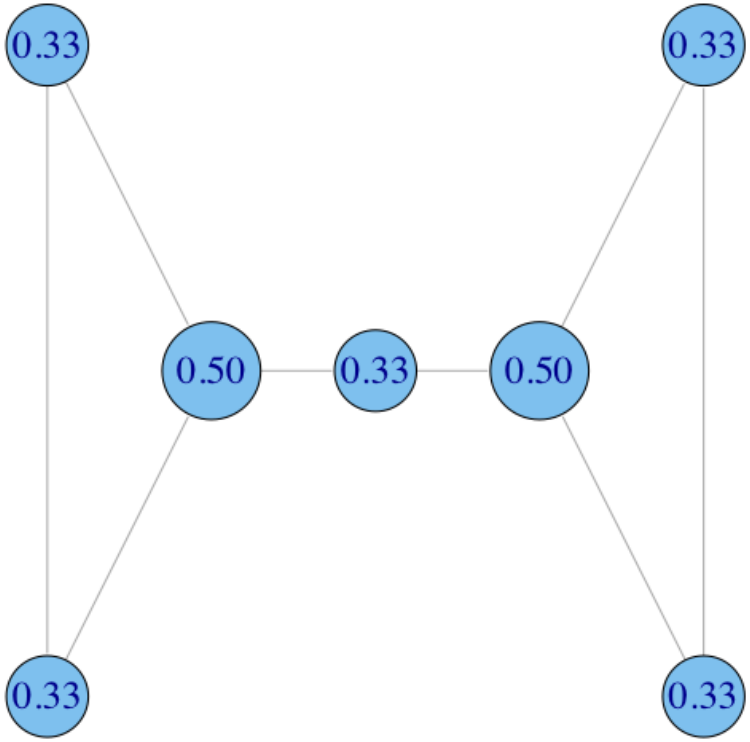
$$C_D(i) = \frac{d(i)}{n-1}.$$

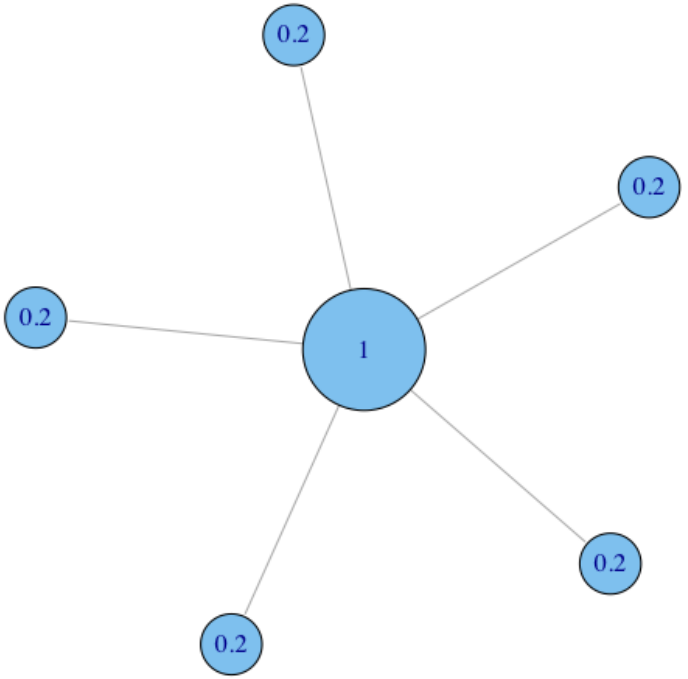
Directed Degree Centrality

- Terhadap Directed Graph:
  - kita perlu membedakan antara aktor in-links *i* (tautan yang menunjuk ke *i*), dan aktor out-links (tautan yang menunjuk keluar dari *i*).
  - Degree centrality didefinisikan berdasarkan hanya pada out-degree (jumlah edge out-links), yaitu $d_o(i)$.

$$C_D'(i) = \frac{d_o(i)}{n-1}.$$

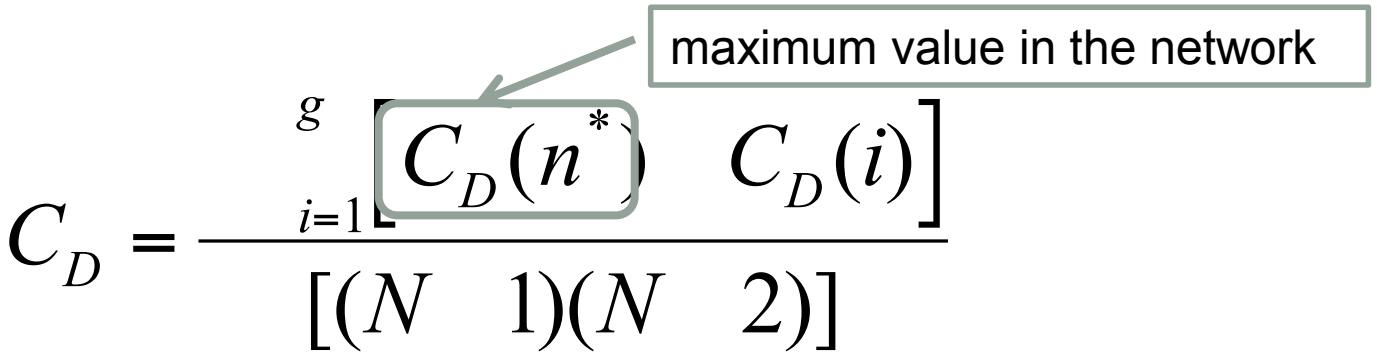
Degree Centrality



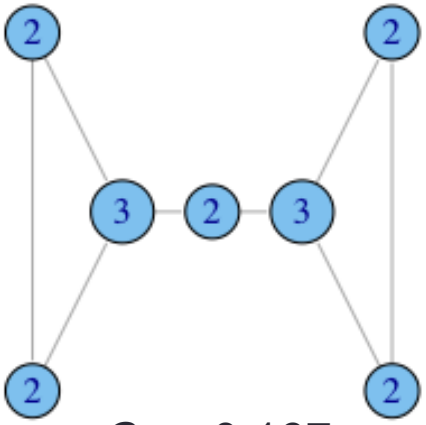


Degree Centrality

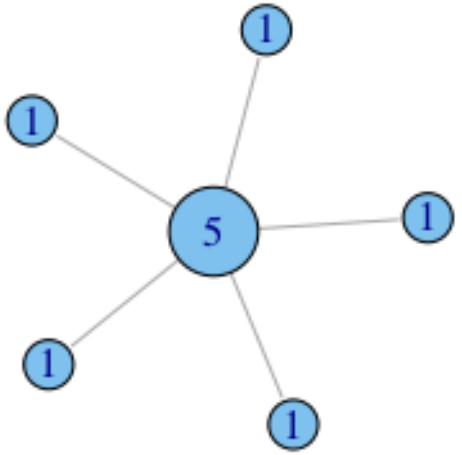
- Berapa banyak variasi yang ada dalam nilai centrality di antara node?
- Rumus Freeman terkait dengan sentralisasi :



Degree Centrality



$C_D = 0.167$



$C_D = 1.0$

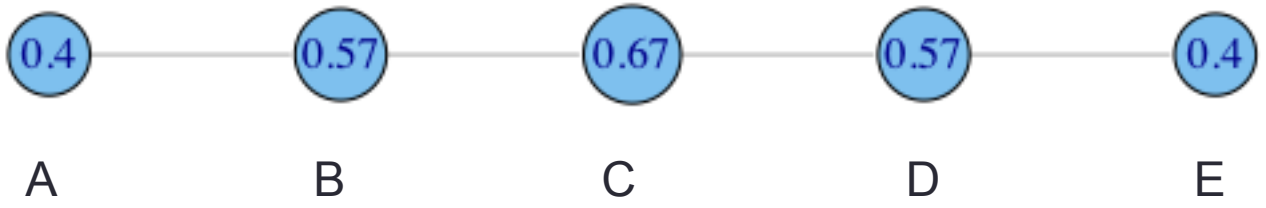
$C_D = 0.167$

Closeness Centrality

- Closeness Centrality didasarkan pada jarak (kedekatan).
- Ide dasarnya bahwa seorang aktif x_i dikatakan sebagai pusat jika aktor tersebut dapat berinteraksi dengan aktor lain secara mudah.
  - yaitu, jarak dari aktor *i* ke aktor lain adalah terpendek.
- Kita dapat menggunakan shortest distance untuk menghitung pengukuran ini.
- Misalkan jarak terpendek dari aktor i ke aktor j adalah d(i,j) (diukur sebagai jumlah tautan dalam sebuah jalur terpendek).

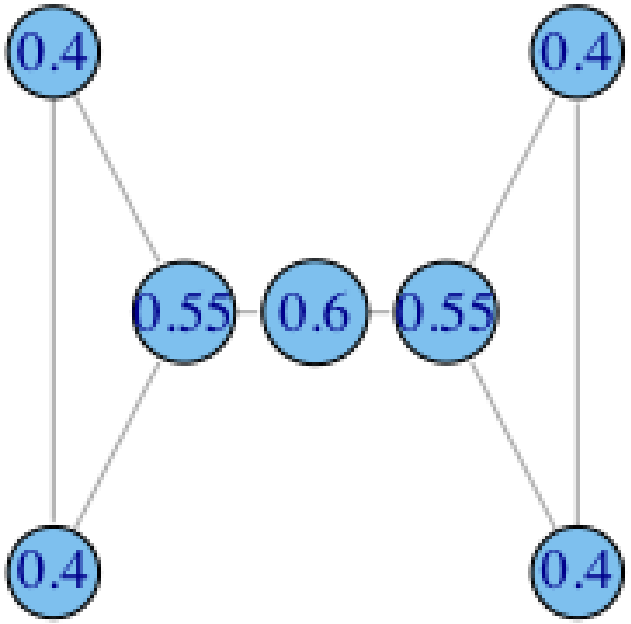
$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}.$$

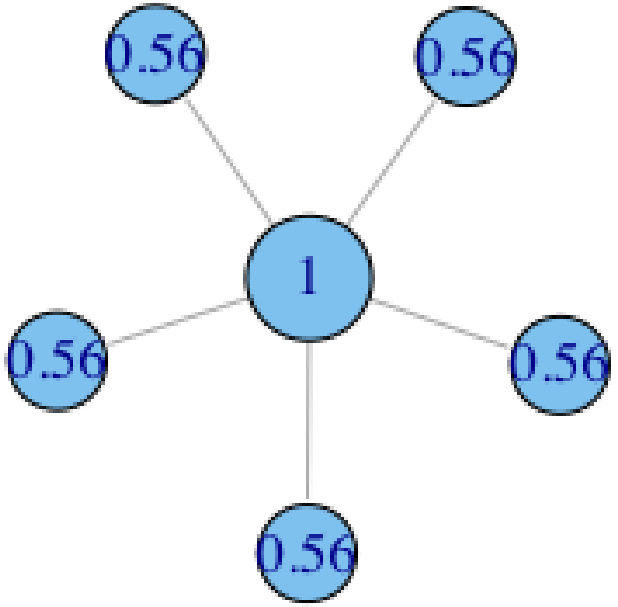
Closeness Centrality



$$C_c'(A) = \frac{N}{\sum_{j=1} d(A,j)} \quad 1 = \frac{N \quad 1}{1+2+3+4} \quad 1 = \frac{10}{4} \quad 1 = 0.4$$

Closeness Centrality





Closeness Centrality

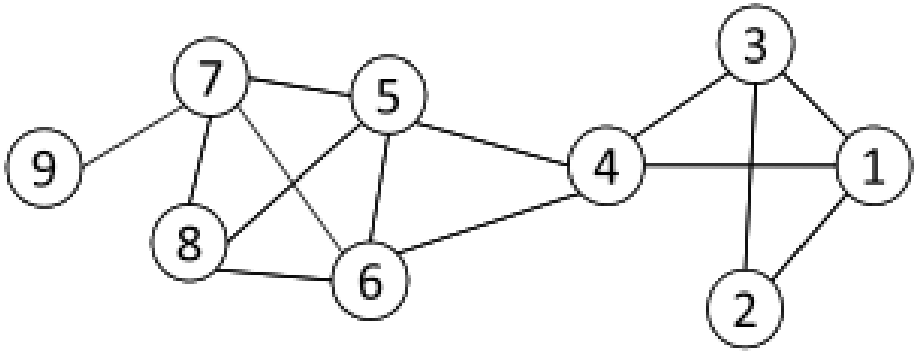


Table 2.1: Pairwise geodesic distance

Node	1	2	3	4	5	6	7	8	9
1	0	1	1	1	2	2	3	3	4
2	1	0	1	2	3	3	4	4	5
3	1	1	0	1	2	2	3	3	4
4	1	2	1	0	1	1	2	2	3
5	2	3	2	1	0	1	1	1	2
6	2	3	2	1	1	0	1	1	2
7	3	4	3	2	1	1	0	1	1
8	3	4	3	2	1	1	1	0	2
9	4	5	4	3	2	2	1	2	0

$$C_C(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = 8/17 = 0.47,$$

$$C_C(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = 8/13 = 0.62.$$

Betweenness Centrality

- Jika ada dua aktor yang saling berdekatan, yaitu j dan k, ingin beriteraksi dan aktor i berada pada jalur hubungan antara j dan k, maka i memiliki kontrol terhadap interaksi keduanya.
- Betweenness mengukur kontrol tersebut.
- sehingga, jika i berada pada jalur dari beberapa interaksi, maka i adalah sebuah aktor penting.

Betweenness Centrality

- Misalkan p_{jk} adalah jumlah jalur terpendek antara aktor j dan k.
- Betweenness seorang aktor i didefinisikan sebagai jumlah jalur terpendek yang melewati i (dinyatakan dengan $p_{jk}(i)$, $j \neq i$ dan $k \neq i$), dinormalisasikan dengan total jumlah jalur terpendek dari semua pasangan aktor, kecuali i:

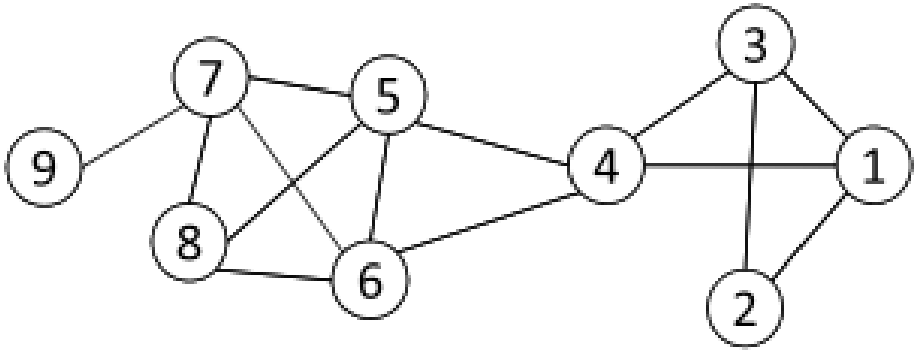
$$C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}.$$

Betweenness Centrality

- mungkin ada beberapa jalur terpendek antara aktor j dan k.
- beberapa jalur tersebut melewati i, dan beberapa jalur lain tidak.
- Kita mengasumsikan bahwa semua jalur digunakan dengan cara yang serupa.
- $C_B(i)$ memiliki nilai minimum 0, yang menyatakan i tidak terletak pada sembarang jalur terpendek.
- $C_B(i)$ memiliki nilai maksimum (n-1)(n-2)/2, yang menunjukkan jumlah pasangan ak
di dalamnya.

$$C_B'(i) = \frac{2\sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}.$$

Betweenness Centrality

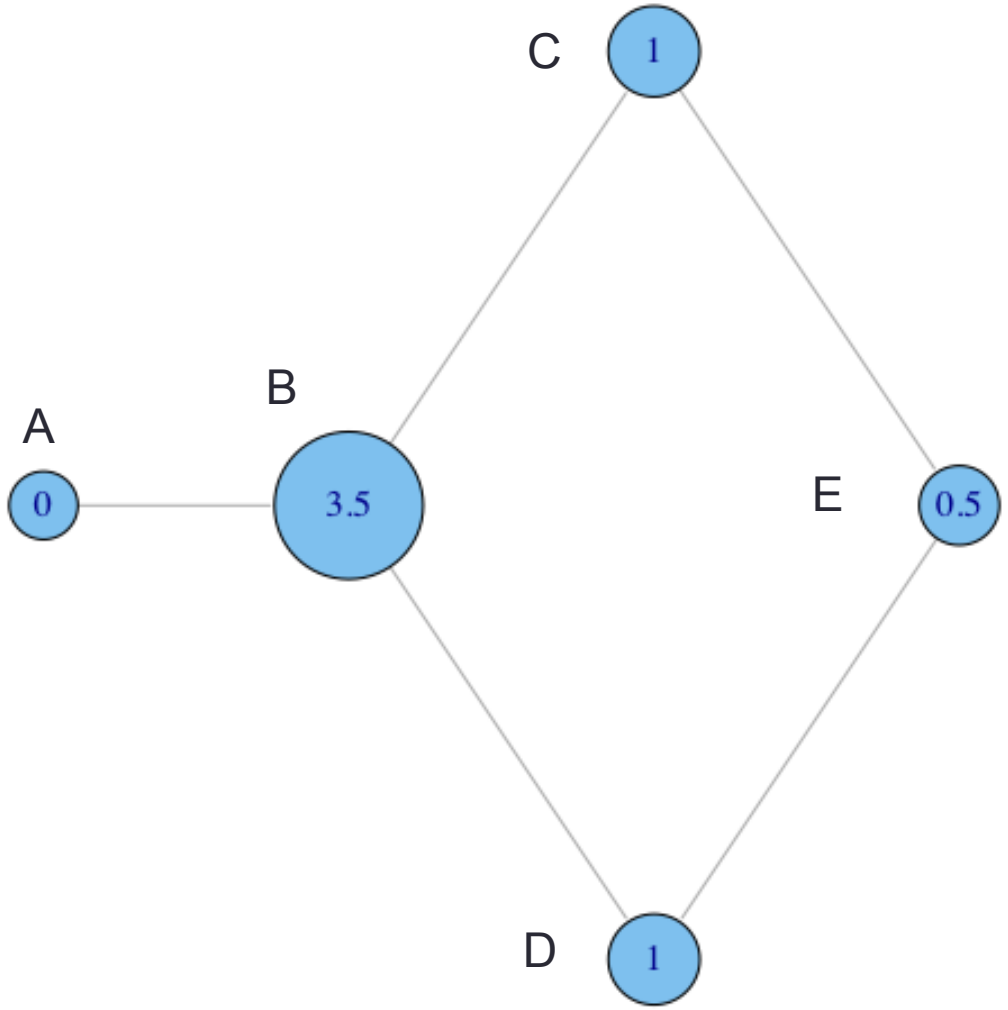


$$C_B(4) = 15$$

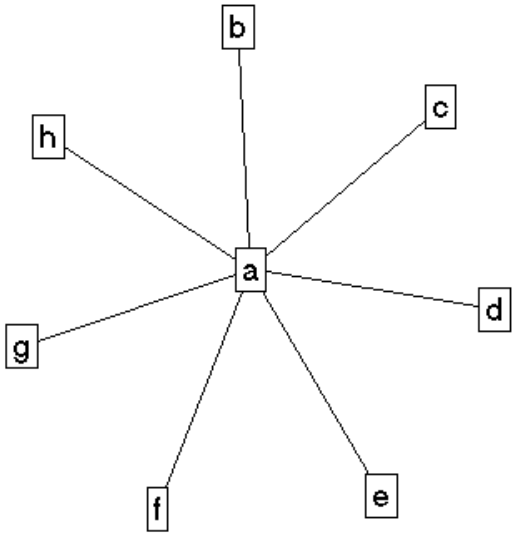
Table 2.2: $\sigma_{st}(4)/\sigma_{st}$

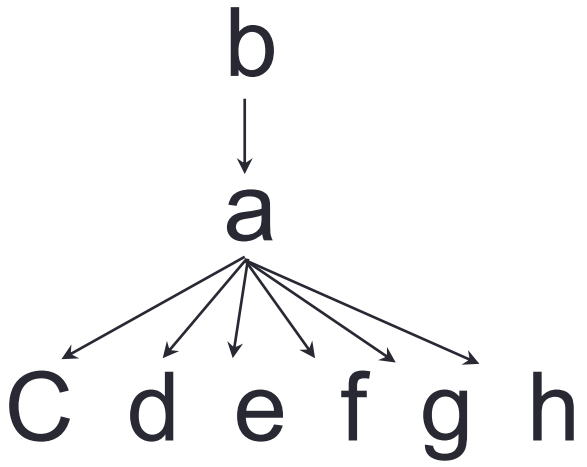
	$s=1$	$s=2$	$s=3$
$t=5$	1/1	2/2	1/1
$t=6$	1/1	2/2	1/1
$t=7$	2/2	4/4	2/2
$t=8$	2/2	4/4	2/2
$t=9$	2/2	4/4	2/2

Betweenness Centrality

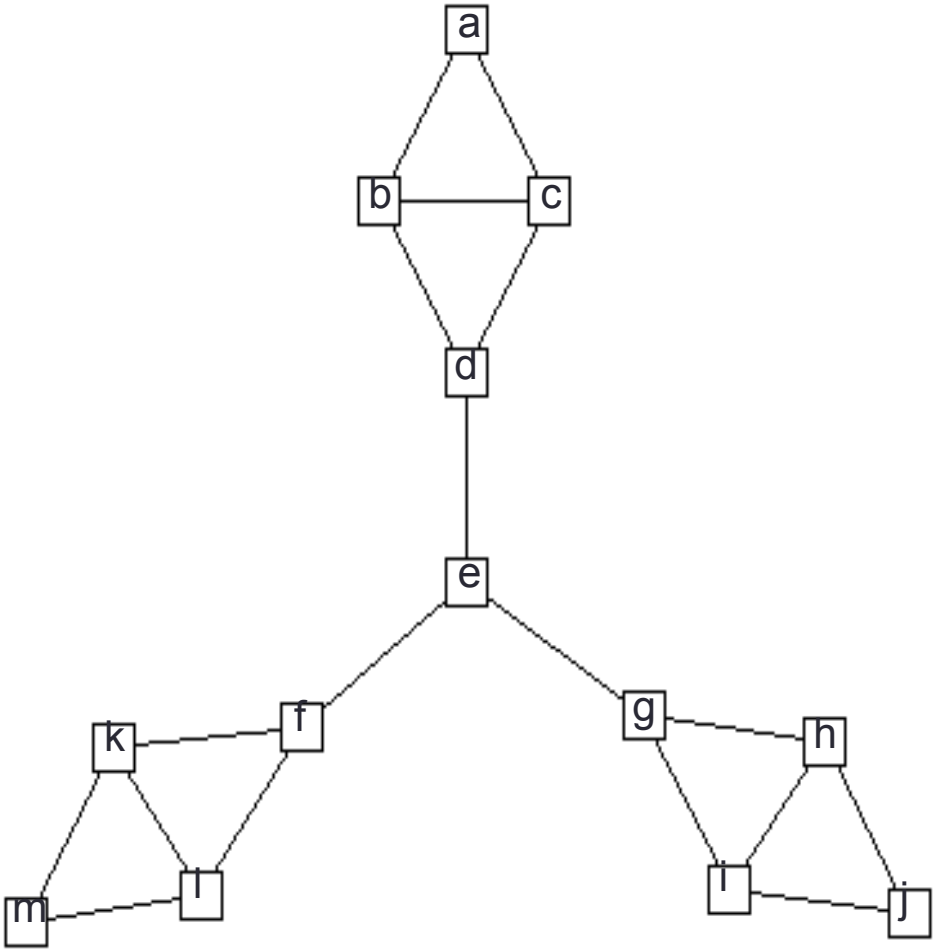


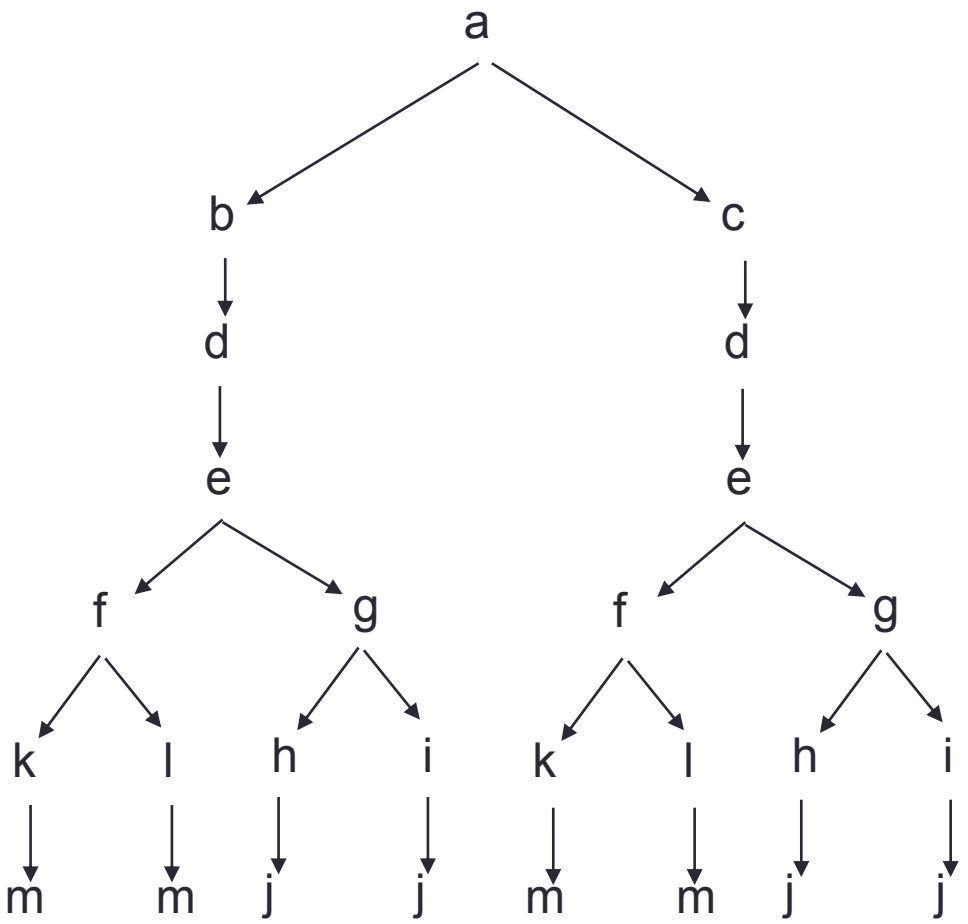
Betweenness Centrality



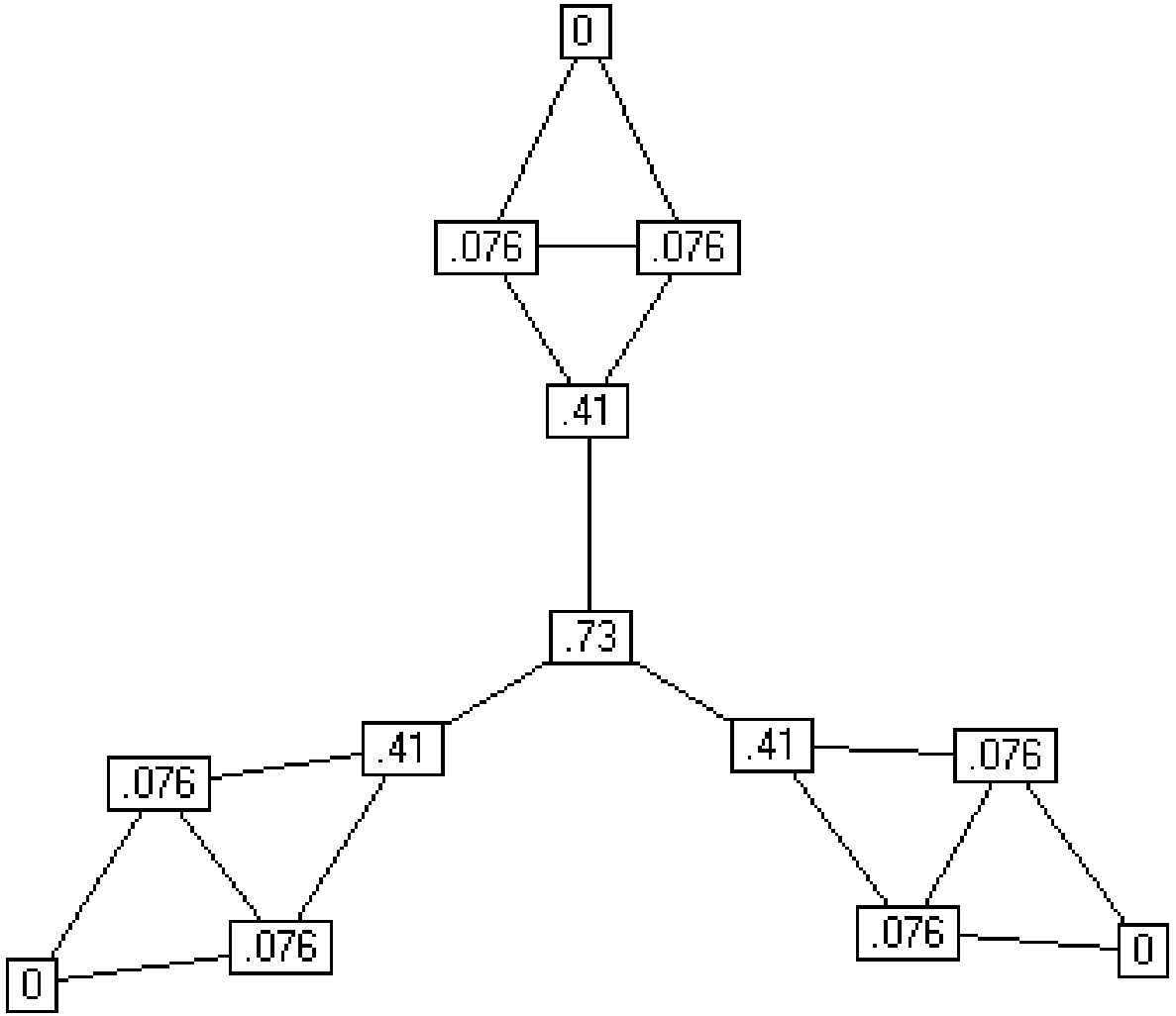


Betweenness Centrality

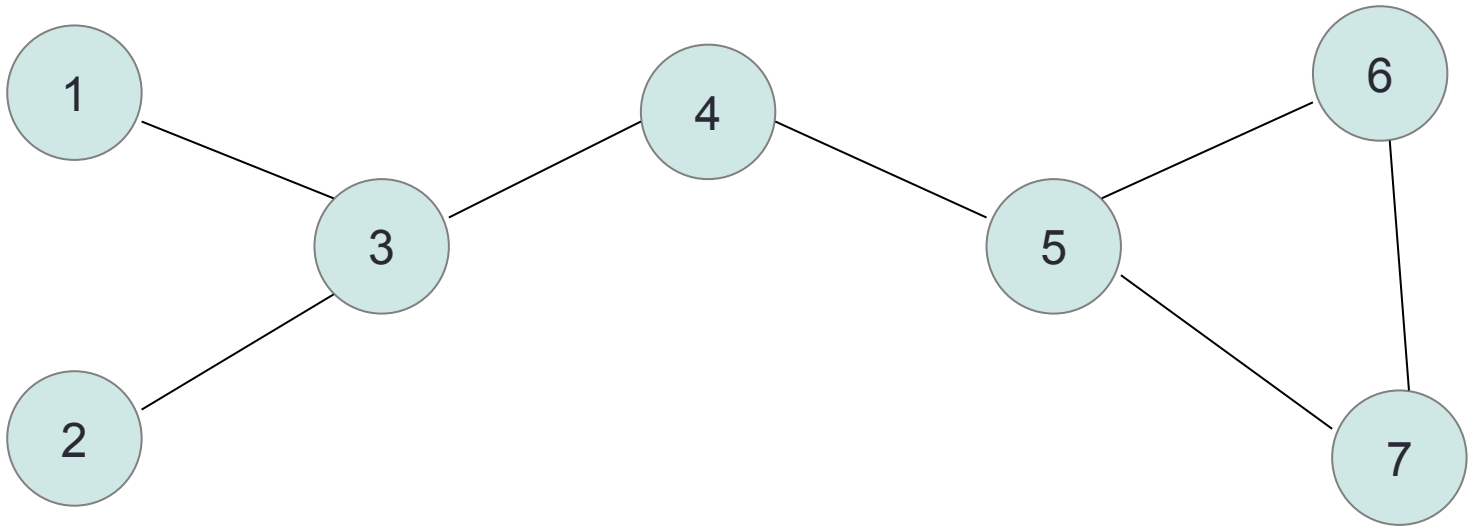




Betweenness Centrality



Contoh



http://www.sscnet.ucla.edu/soc/faculty/mcfarland/soc112/cent-ans.htm

Prestige

- *Prestige* (martabat/wibawa) merupakan suatu pengukuran yang lebih halus terhadap peran seorang aktor daripada pengukuran *centrality*.
- Kita perlu membedakan antara ikatan keluar (*out-links*) ikatan masuk (*in-links*).
- Seorang aktor bermartabat tinggi jika aktor tersebut memiliki ikatan sebagai penerima (*in-links*).
- Perbedaan utama antara konsep *centrality* dan *prestige* adalah *centrality* fokus pada *out-links*, sementara *prestige* fokus pada *in-links*.

Degree Prestige

- Seorang aktor dikatakan *prestigious* jika ia menerima banyak *in-links* atau nomasi.

$$P_D(i) = \frac{d_I(i)}{n-1}$$

- dimana $d_I(i)$ adalah in-degree dari i (jumlah in-links dari i) dan n adalah total jumlah aktor dalam jaringan.

TERIMA KASIH

Budi Susanto

WEB USAGE MINING

Budi Susanto

Web Mining

- Web mining adalah aplikasi teknik data mining untuk menyarikan pengetahuan dari data Web.
- Data web adalah
  - web content
    - text, image, records, dsb.
  - web structure
    - hyperlinks, tags, dsb.
  - web usage
    - log httpd, log app server, dsb.

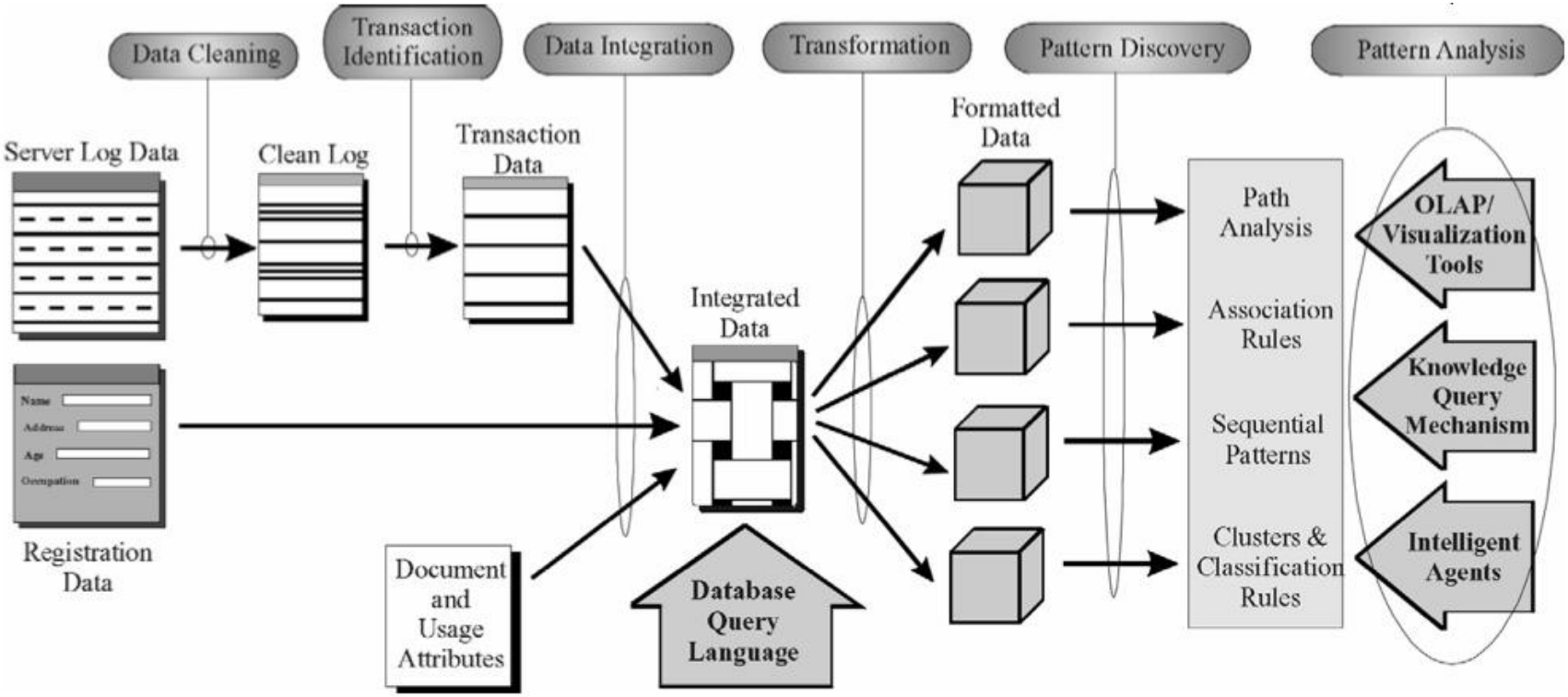
Preprocessing Web Data

- Web Content
  - menyarikan “potongan” dari sebuah dokumen Web
  - Metode yang digunakan Information Retrieval, Klasifikasi, Clustering.
- Web Structure
  - mengidentifikasikan pola-pola graf menarik tertentu bersama suatu metric
  - Analisis hyperlink: PageRank, HITS, SNA
- Web Usage
  - identifikasi user, pembuatan sesi, pendeteksian dan penyaringan robot, menyarikan pola pemakaian.

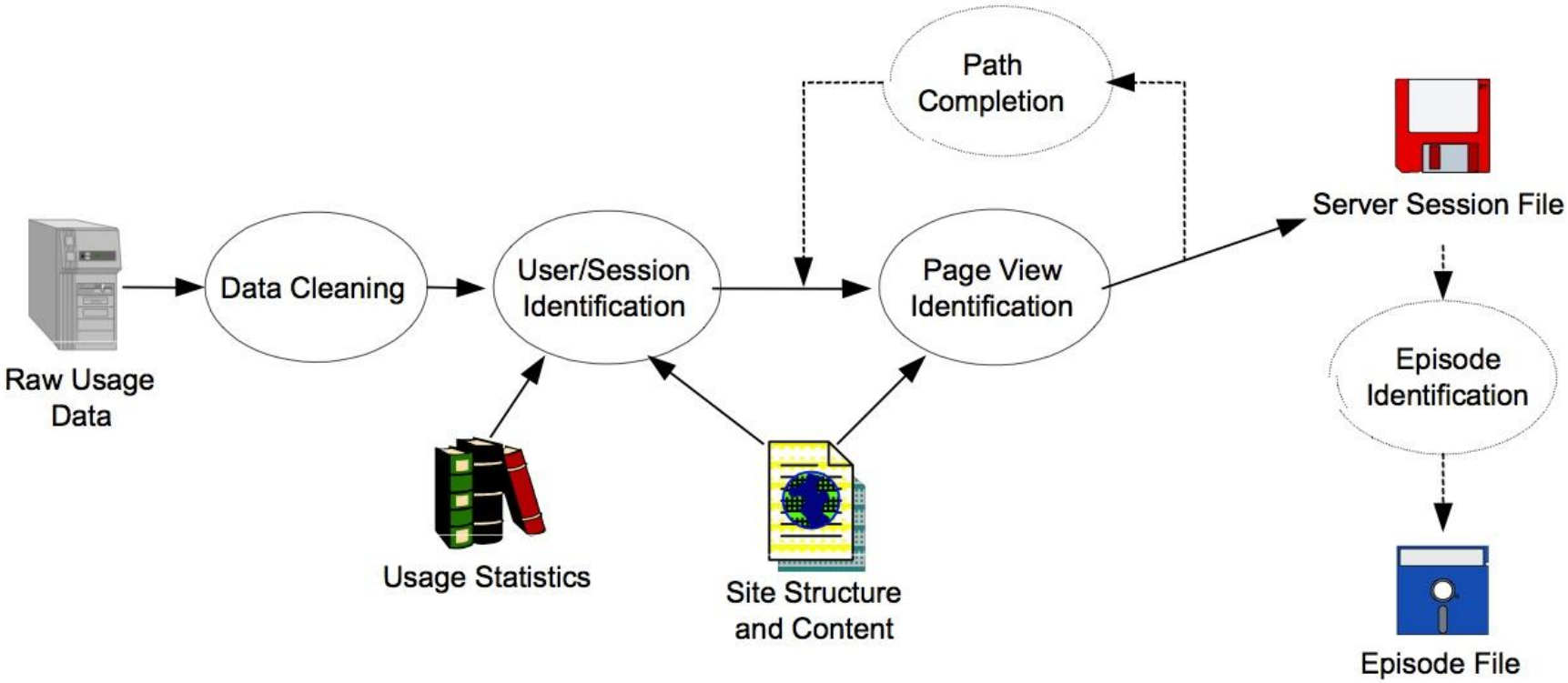
Web Usage Mining

- Sebuah web adalah sekumpulan inter-related file pada satu atau lebih web server
- Web Usage Mining
  - Menemukan pola dari data yang dihasilkan oleh transaksi client-server pada satu atau lebih web server
- Sumber data
  - data yang dihasilkan otomatis oleh server dalam bentuk access log, referrer log, agent log, client-side cookie
  - user profile
  - meta data: atribut halaman, atribut content, usage data

Web Usage Mining Process



Arsitektur Preprocessing



Format Log NCSA

- Log yang dihasilkan web server yang mencatat "*what happened when by whom*".
- Contoh:

```
127.0.0.1 user-identifier frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326
```

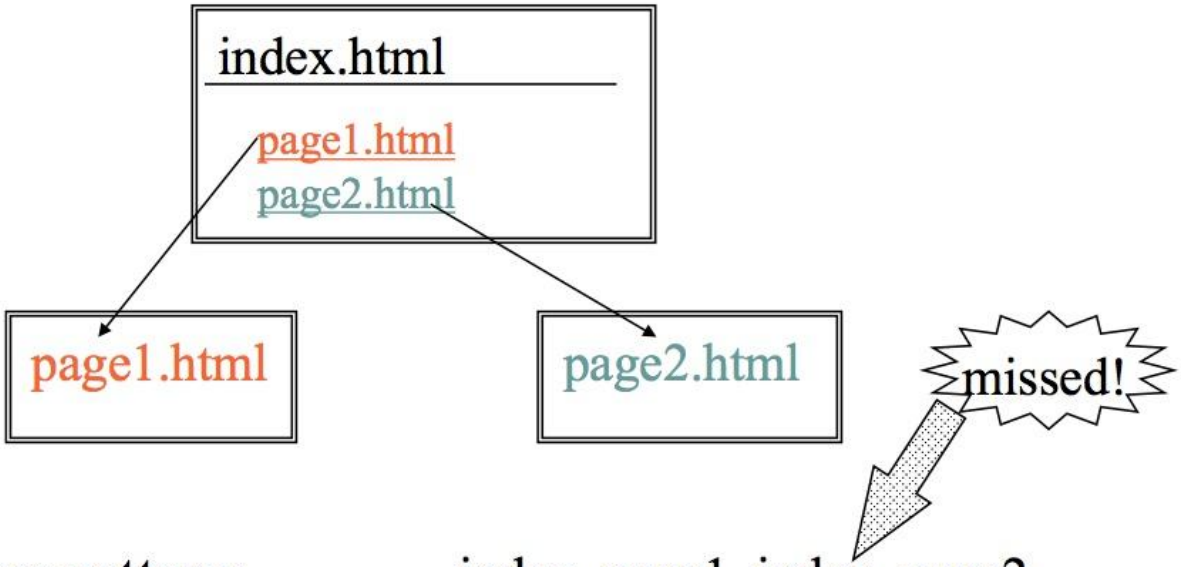
```
repo.ukdw.ac.id 173.199.114.211 - - [05/May/2013:10:28:04 +0700] "GET /slackware/slackware-14.0/slackware/x/libdrm-2.4.33-i486-1.txz.asc HTTP/1.1" 200 198
ti.ukdw.ac.id 66.249.74.54 - - [05/May/2013:10:28:09 +0700] "GET /images/events/sony_step_up.jpg HTTP/1.1" 304 -
skripsi.ukdw.ac.id 103.10.64.7 - - [05/May/2013:10:28:09 +0700] "GET / HTTP/1.1" 200 13814
repo.ukdw.ac.id 180.76.5.62 - - [05/May/2013:10:28:10 +0700] "GET /ubuntu/pool/main/c/command-not-found/?C=D;O=D HTTP/1.1" 200 2557
repo.ukdw.ac.id 180.76.5.189 - - [05/May/2013:10:28:20 +0700] "GET /slackware/slackware64-current/testing/packages/xorg-server-1.14.x/xf86-video-v4l-0.2.0-x86
_64-7.txz.asc HTTP/1.1" 200 198
scripti.ukdw.ac.id 114.79.16.8 - - [05/May/2013:10:28:26 +0700] "GET / HTTP/1.1" 200 12558
scripti.ukdw.ac.id 114.79.16.8 - - [05/May/2013:10:28:27 +0700] "GET /css/form.cake.generic.css HTTP/1.1" 200 9266
```

Persoalan Usage Data

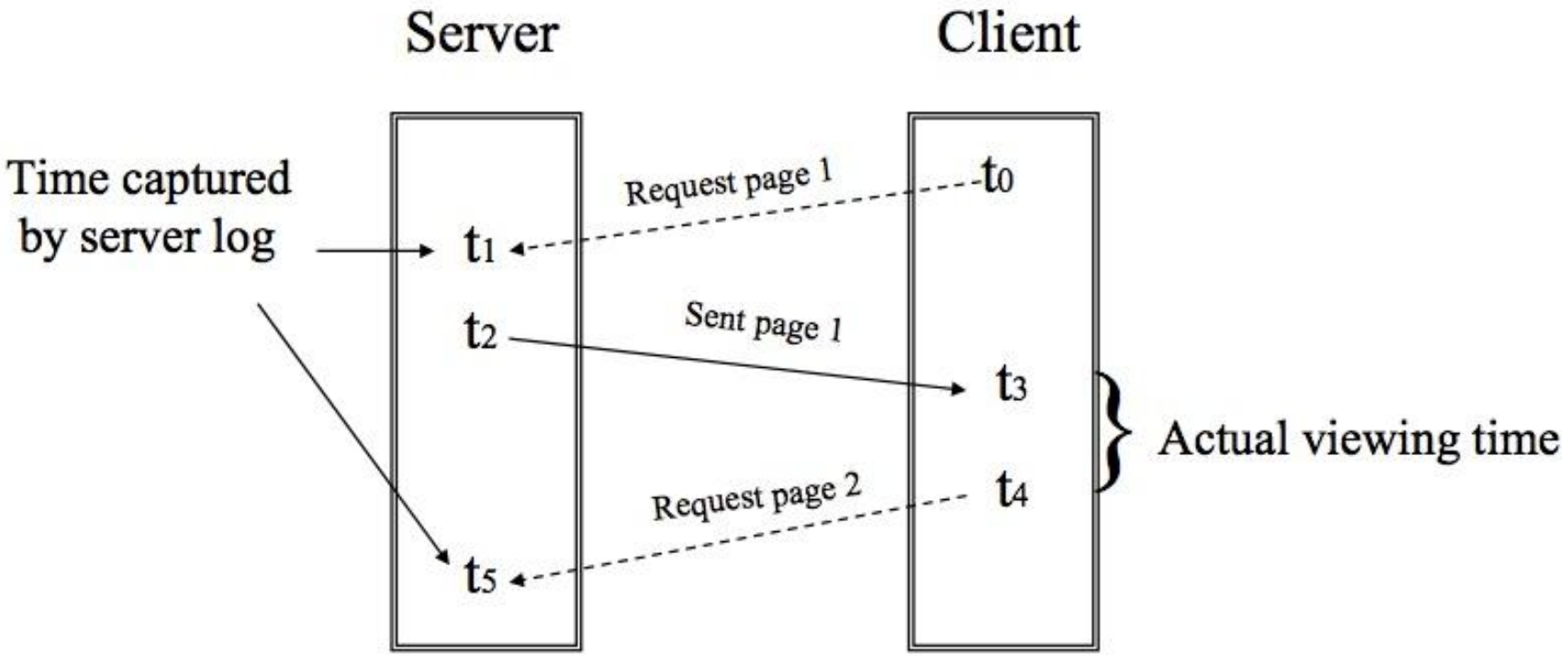
- Pengenalan terhadap Session
  - Cookie, User Login, SessionID, IP+Agent, Client-side tracking
- Data CGI
  - GET dan POST
- Caching
- Dynamic Page
- Deteksi Robot dan Penyaringan
- Pengenalan Transaksi
  - mengenal user
  - mengenal transaksi user

Masalah terhadap Caching

- Client dan proxy server menyimpan local copy secara lokal
- pemakaian tombol “Back” atau “Forward” pada browser, akan mengakses local copy daripada mengakses web server kembali.

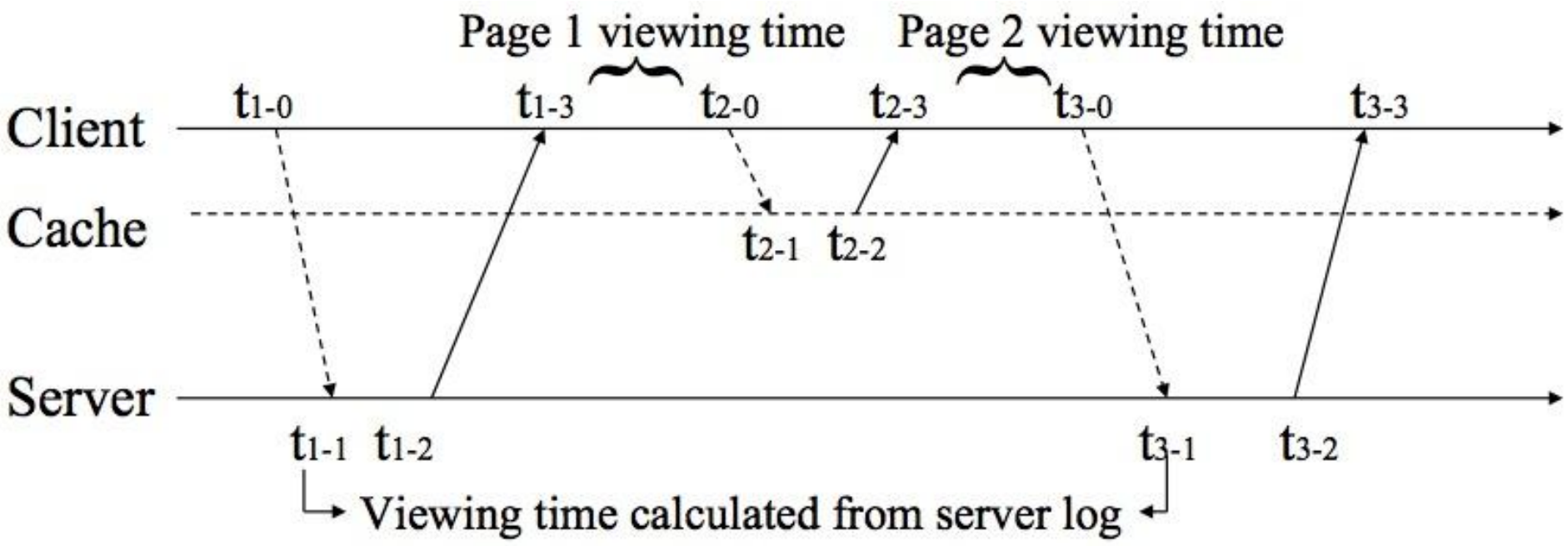


Kesalahan Penyimpanan Waktu Akses



Kehilangan Page View di Server

- Contoh urutan waktu akses yang hilang karena adanya proxy



Deteksi Robot

- Robot Web adalah program yang secara otomatis menjelajah struktur hyperlink dari WWW dalam rangka untuk mendapatkan lokasi dan mengambil informasi.
- Motivasi adalah membedakan mana yang robot dan mana yang diakses dari user.

Identifikasi Transaksi

- Pertanyaan utama:
  - bagaimana mengenal pemakai unik
  - bagaimana mendefinisikan transaksi seorang user
- Masalah-masalah
  - alamat IP komputer tunggal akan tersembunyikan dibalik proxy server
  - client-side dan proxy caching membuat server log kurang handal
  - user id biasanya disembunyikan terkait dengan keamanan
- Solusi standar
  - registrasi pemakai
  - client-side cookie
  - *cache busting*

Identifikasi Transaksi

- Mengenal User Session
  - menggunakan field IP, Agent, dan OS sebagai atribut kunci
  - menggunakan client-side cookie dan user id unik (jika tersedia)
  - menggunakan session time-out
  - menggunakan sinkronisasi log dan timestamp untuk memperluas user path dari sebuah session
  - memanfaatkan atribut halaman (ukuran, tipe), panjang reference

Analisis Transaksi Web

- Association Rule
- Sequential Pattern
- Clustering dan Classification

TERIMA KASIH