Text Mining

Intro to Web Mining & SNA

| Outline

1. Intro to Web Mining
2. Web Usage Mining
3. Social Network Analysis
4. Centrality
5. SNA example

| #1 Intro to Web Mining

- Web mining adalah aplikasi teknik data mining untuk mencari pengetahuan/knowledge 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.

| #1 Intro to Web Mining

Web Mining bertujuan untuk menemukan informasi atau pengetahuan dari:

- Web hyperlink structure
  - menemukan halaman web terpenting
  - menemukan komunitas pemakai yang berbagi ketertarikan topik yang sama
- Page content
  - Ekstraksi data/informasi dari halaman web
  - Integrasi dan pencocokan skema informasi beberapa web
  - Ekstraksi opini
  - Knowledge synthesis
  - Segmentasi halaman web dan mendeteksi noise
- Usage data.
  - menemukan pola akses pemakai terhadap web, melalaui click stream.

| #1 Intro to Web Mining

Karakteristik data web:

- jumlah data/informasi di web sangat besar dan terus bertambah.
- tipe data beragam.
- informasi pada web sangat beragam.
- informasi-informasi di web saling terhubung.
- informasi di web sangat "kotor“.
- web juga merupakan service.
- web dinamis.
- web merupakan sarana komunitas sosial virtual.

| #1 Intro to Web Mining

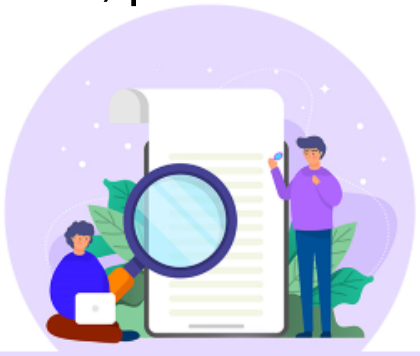
Web Mining task:

1. Web Structure Mining
   - Wrapper induction
   - Structure Matching: edit distance
2. Web Content Mining
   - Machine learning dan information extraction
3. Web Usage Mining
   - Sequence pattern analysis

| #1 Intro to Web Mining

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.



| #2 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

| #2 Web Usage Mining

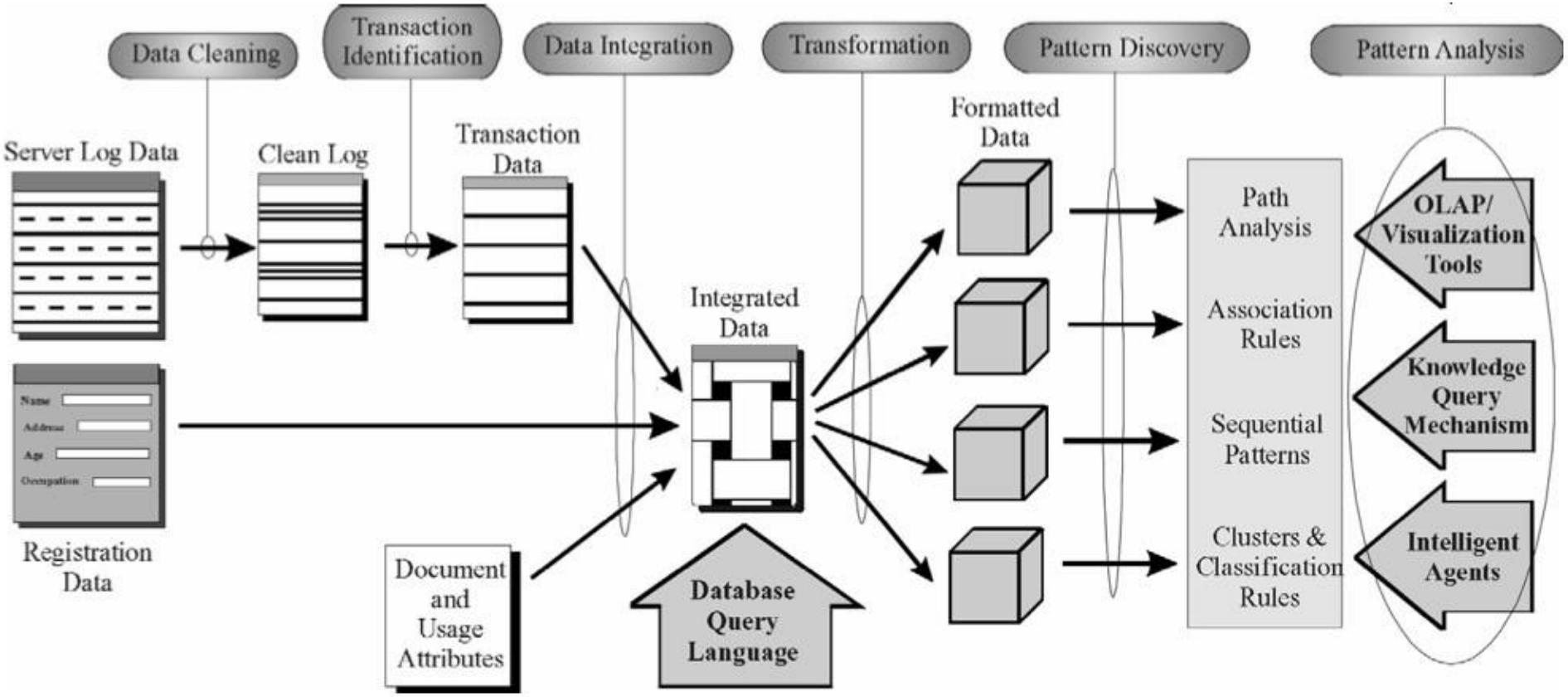
Format Log NCSA

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

Contoh:

```
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
128.159.129.170 - - [01/Aug/1995:00:13:16 -0400] "GET / HTTP/1.0" 200 7280
kgtyk4.kj.yamagata-u.ac.jp - - [01/Aug/1995:00:00:17 -0400] "GET / HTTP/1.0" 200 7280
rpgopher.aist.go.jp - - [01/Aug/1995:00:01:58 -0400] "GET /ksc.html HTTP/1.0" 200 7280
204.248.155.42 - - [01/Aug/1995:00:05:32 -0400] "GET /icons/menu.xbm HTTP/1.0" 200 527
204.248.155.42 - - [01/Aug/1995:00:05:32 -0400] "GET /icons/image.xbm HTTP/1.0" 200 509
143.158.26.50 - - [01/Aug/1995:00:14:07 -0400] "GET / HTTP/1.0" 200 7280
ai.asu.edu - - [01/Aug/1995:00:03:55 -0400] "GET /facts/faq01.html HTTP/1.0" 200 19320
gw1.att.com - - [01/Aug/1995:00:03:56 -0400] "GET /icons/menu.xbm HTTP/1.0" 304 0
gw1.att.com - - [01/Aug/1995:00:03:56 -0400] "GET /icons/text.xbm HTTP/1.0" 304 0
143.158.26.50 - - [01/Aug/1995:00:14:07 -0400] "GET / HTTP/1.0" 200 7280
```

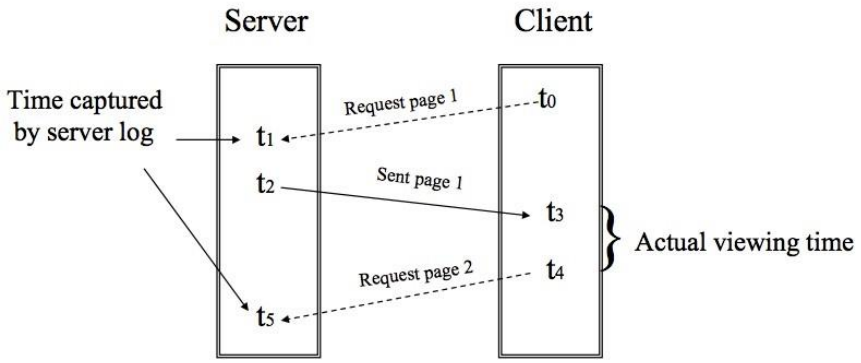
| #2 Web Usage Mining (process)



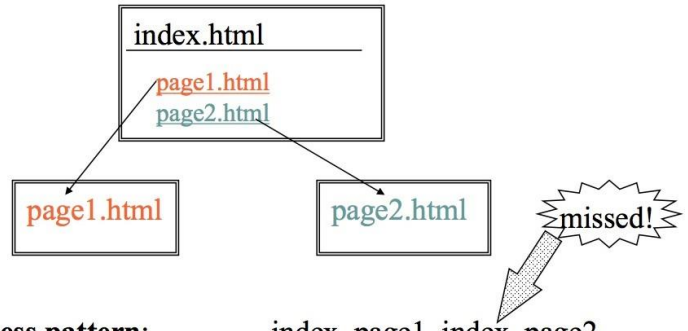
| #2 Web Usage Mining

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
- Penyimpanan Waktu



- 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.

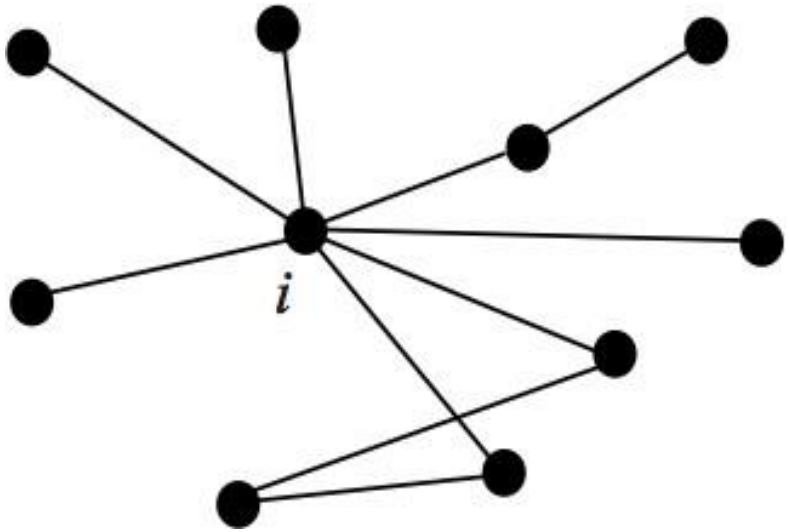


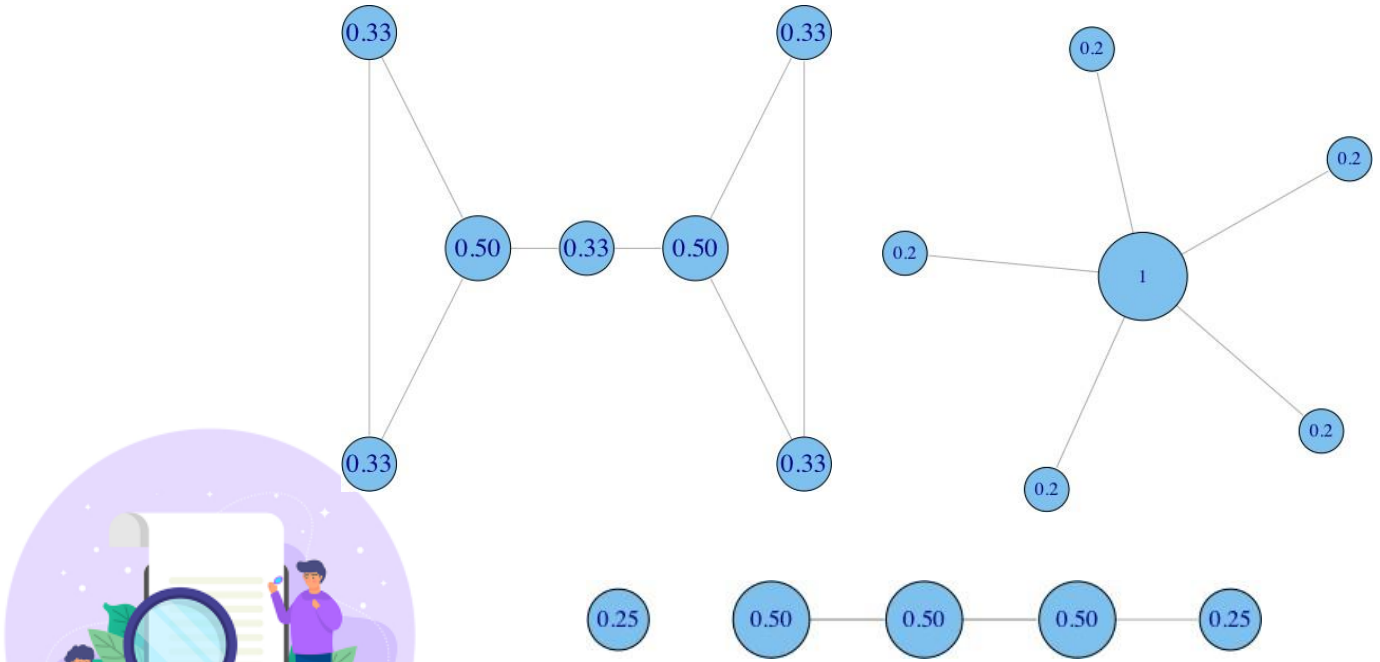
| #3 Social Network Analysis

- Social network adalah studi terhadap entitas sosial (misalnya orang dalam suatu organisasi), dan interaksi serta relasi antar 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.

#4 Centrality

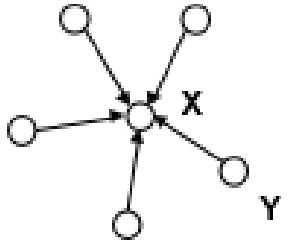
- Dalam kontek suatu organisasi, seseorang dengan hubungan atau komunikasi yang ekstensif dengan banyak orang lain dalam organisasi dinilai lebih penting daripada orang lain yang memiliki kontak lebih sedikit
- Tautan atau hubungan dapat juga disebut sebagai ikatan (*ties*).
- Seorang aktor pusat terlibat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



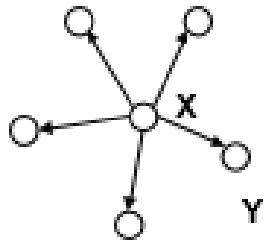


#4 Centrality (con't)

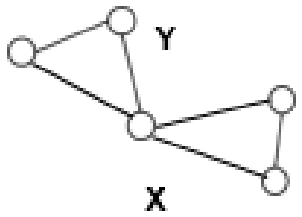
- Degree centrality
  - out-links
  - in-links
- Closeness centrality
- Betweeness centrality



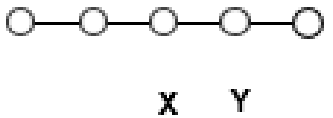
indegree



outdegree



betweenness

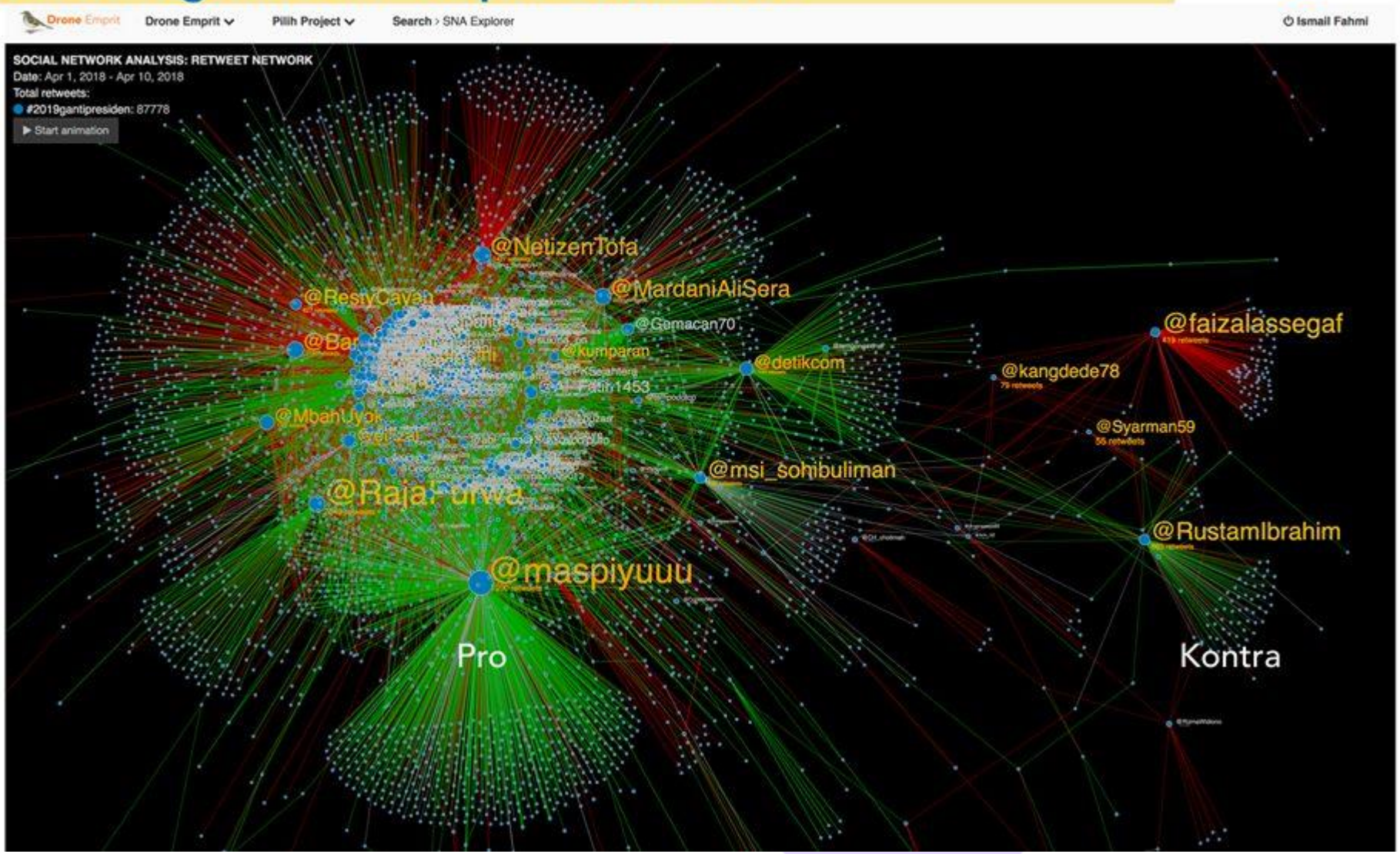


closeness

#5 SNA (example)

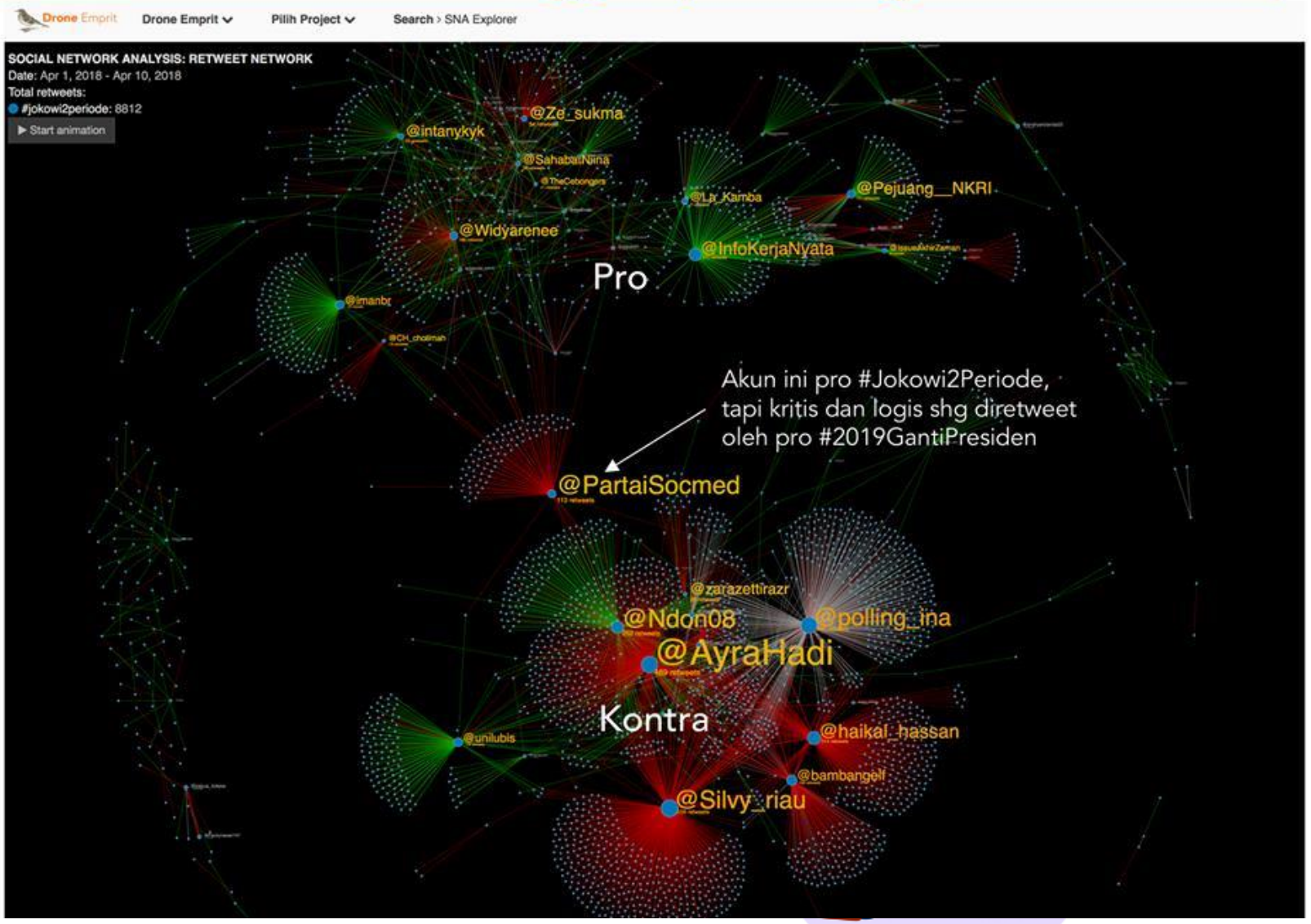
SNA #2019GantiPresiden

Akun Organik dan Tanpa Perlawanan



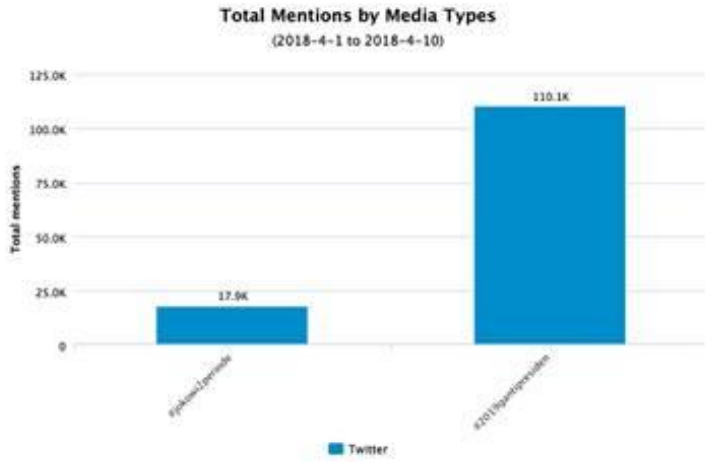
#5 SNA (example)

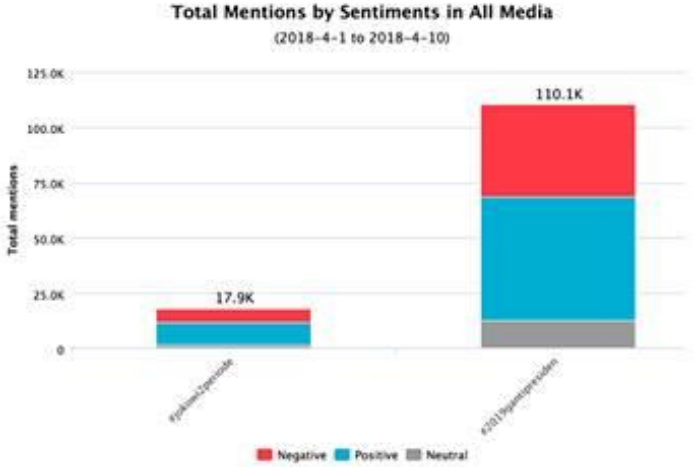
SNA #Jokowi2Periode
Cluster Pro belum menggarap hashtag ini?

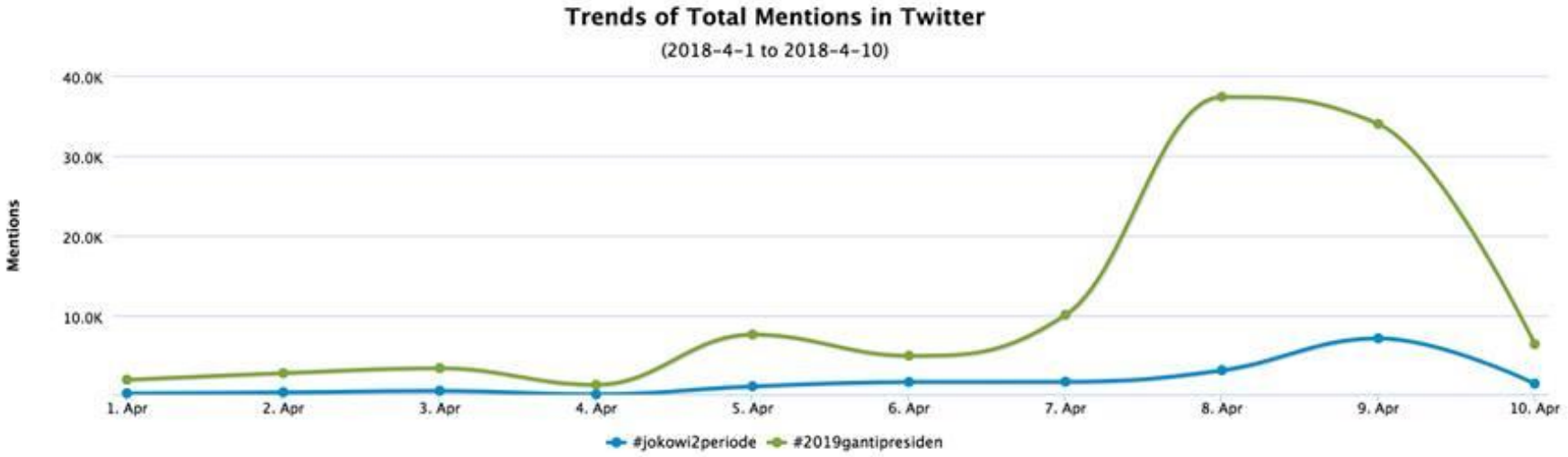


#5 SNA (example)

Volume dan Tren
#Jokowi2Periode vs #2019GantiPresiden







Thank you

</ti>
FST | UINAM
teknik informatika

UNIVERSITAS ISLAM NEGERI
ALAUDDIN
MAKASSAR