# Social Media Analytic Workshop

Digital Business Ecosystem Research Center

Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# 1

# Handshake with Social Media Analytic

Social Media Analytic Workshop

Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Learning Outline

- **Social Media Analytic Definition**
- Background
- The Benefits of Social Media Analytic
- Challenges and Opportunity
- Understanding Data
- Social Media Analytic Workflow
- Social Media Analytic Implementation

# Social Media Analytic Definition

- Social media analytics (SMA) refers to the approach of collecting data from social media sites and blogs and evaluating that data to make insightful decision making.
- It can also analyze online media channels such as news websites, blogs, and forums.

# Learning Outline

- Social Media Analytic Definition
- **Background**
- The Benefits of Social Media Analytic
- Challenges and Opportunity
- Understanding Data
- Social Media Analytic Workflow
- Social Media Analytic Implementation

# Background

- Active Social Media User
- Time Spent with Social Media
- Human Data Production

JAN 2019

# INDONESIA

THE ESSENTIAL HEADLINE DATA YOU NEED TO UNDERSTAND MOBILE, INTERNET, AND SOCIAL MEDIA USE

| TOTAL POPULATION | MOBILE SUBSCRIPTIONS | INTERNET USERS | ACTIVE SOCIAL MEDIA USERS | MOBILE SOCIAL MEDIA USERS |
|---|---|---|---|---|
| 268.2 MILLION | 355.5 MILLION | 150.0 MILLION | 150.0 MILLION | 130.0 MILLION |
| URBANISATION: | vs. POPULATION: | PENETRATION: | PENETRATION: | PENETRATION: |
| 56% | 133% | 56% | 56% | 48% |



**SOURCES:** POPULATION: UNITED NATIONS; U.S. CENSUS BUREAU. MOBILE: GSMA INTELLIGENCE. INTERNET: INTERNETWORLDSTATS; ITU; WORLD BANK; CIA WORLD FACTBOOK; EUROSTAT; LOCAL GOVERNMENT BODIES AND REGULATORY AUTHORITIES; MIDEASTMEDIA.ORG; REPORTS IN REPUTABLE MEDIA. SOCIAL MEDIA: PLATFORMS' SELF-SERVE ADVERTISING TOOLS; PRESS RELEASES AND INVESTOR EARNINGS ANNOUNCEMENTS; ARAB SOCIAL MEDIA REPORT; TECHRASA; NIKI AGHAEI; ROSE.RU. (ALL LATEST AVAILABLE DATA IN JANUARY 2019).

JAN 2019

# TIME SPENT WITH MEDIA

AVERAGE DAILY TIME SPENT CONSUMING AND INTERACTING WITH MEDIA [SURVEY BASED]

AVERAGE DAILY TIME SPENT USING THE INTERNET VIA ANY DEVICE

**8H 36M**

AVERAGE DAILY TIME SPENT USING SOCIAL MEDIA VIA ANY DEVICE

**3H 26M**

AVERAGE DAILY TV VIEWING TIME (BROADCAST, STREAMING AND VIDEO ON DEMAND)

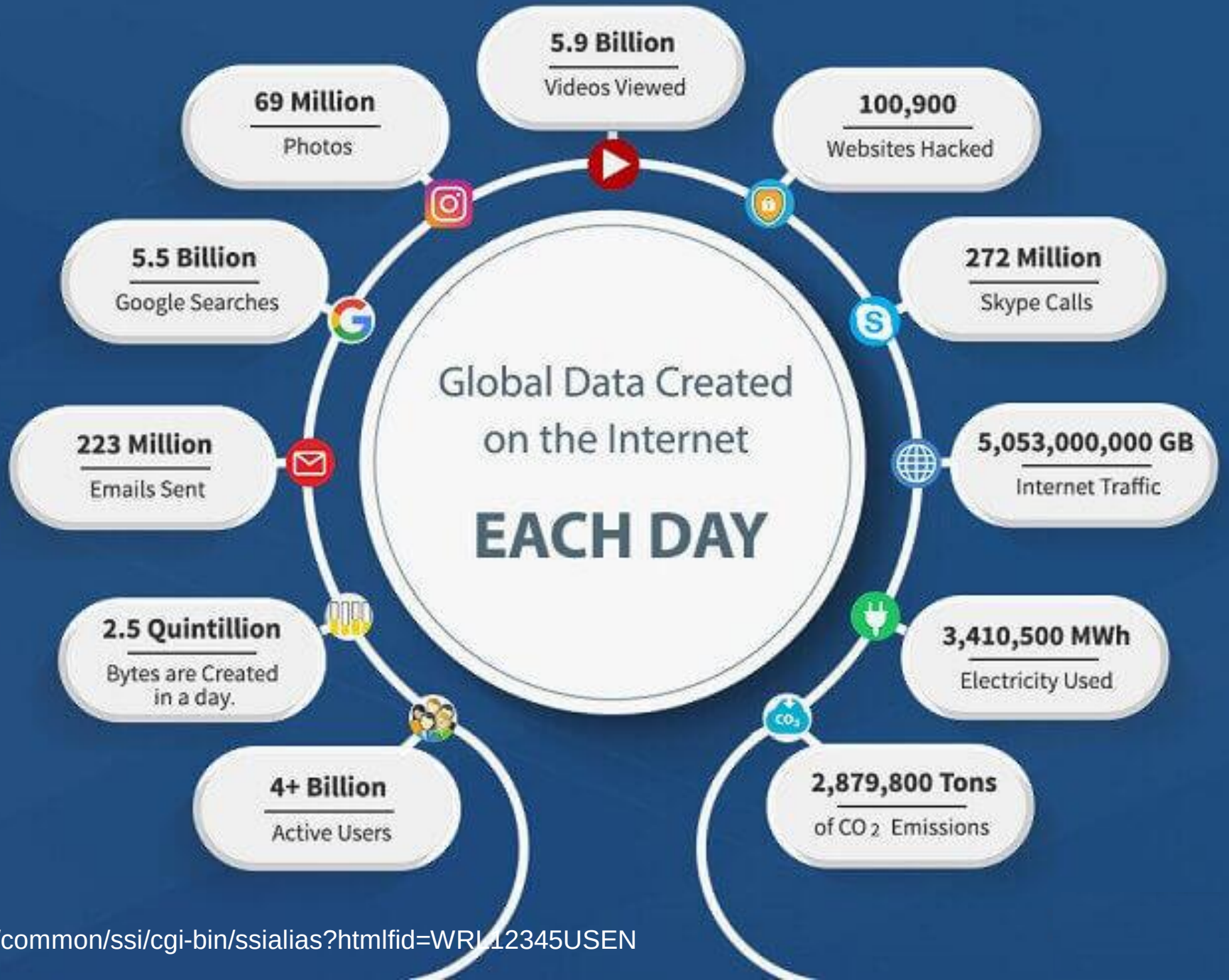
**2H 52M**

AVERAGE DAILY TIME SPENT LISTENING TO STREAMING MUSIC

**1H 22M**



**SOURCE:** GLOBALWEBINDEX (Q2 & Q3 2018). FIGURES REPRESENT THE FINDINGS OF A BROAD SURVEY OF INTERNET USERS AGED 16-64. TIMES ARE DAILY AVERAGES, REPORTED IN HOURS AND MINUTES. **NOTE:** SOME ACTIVITIES MAY TAKE PLACE CONCURRENTLY.




https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN

# Learning Outline

- Social Media Analytic Definition
- Background
- **The Benefits of Social Media Analytic**
- Challenges and Opportunity
- Understanding Data
- Social Media Analytic Workflow
- Social Media Analytic Implementation

# The Benefits of Social Media Analytic

- Create winning social media campaigns
- Find the best influencers for your brand
- Benchmark against your competitors
- Identify trending topics

# Learning Outline

- Social Media Analytic Definition
- Background
- The Benefits of Social Media Analytic
- **Challenges and Opportunities**
- Understanding Data
- Social Media Analytic Workflow
- Social Media Analytic Implementation

# Challenges

- Huge volumes of data
- Incomplete pictures
- Data relevance
- Data quality

# Opportunities

- Uncovering hidden truth
- Track social campaigns
- Create better content
- Competitive benchmarking

# Learning Outline

- Background
- Social Media Analytic Definition
- The Benefits of Social Media Analytic
- Challenges and Opportunities
- **Understanding Data**
- Social Media Analytic Workflow
- Social Media Analytic Implementation

# Understanding Data

## Structured Data

High Degree of Organization, such as a relational database

| Column | Value |
|---|---|
| Patient | John Brown |
| Date of Birth | 12/07/1993 |
| Date Admitted | 02/03/2011 |

## Unstructured Data

Information that is difficult to organize using traditional mechanisms

"The patient came in complaining of chest pain, shortness of breath, and lingering headaches.. Smokes 2 packs a day.. Family history of heart disease.. Has been experiencing similar symptoms for the past 12 hours…"

# Understanding Data

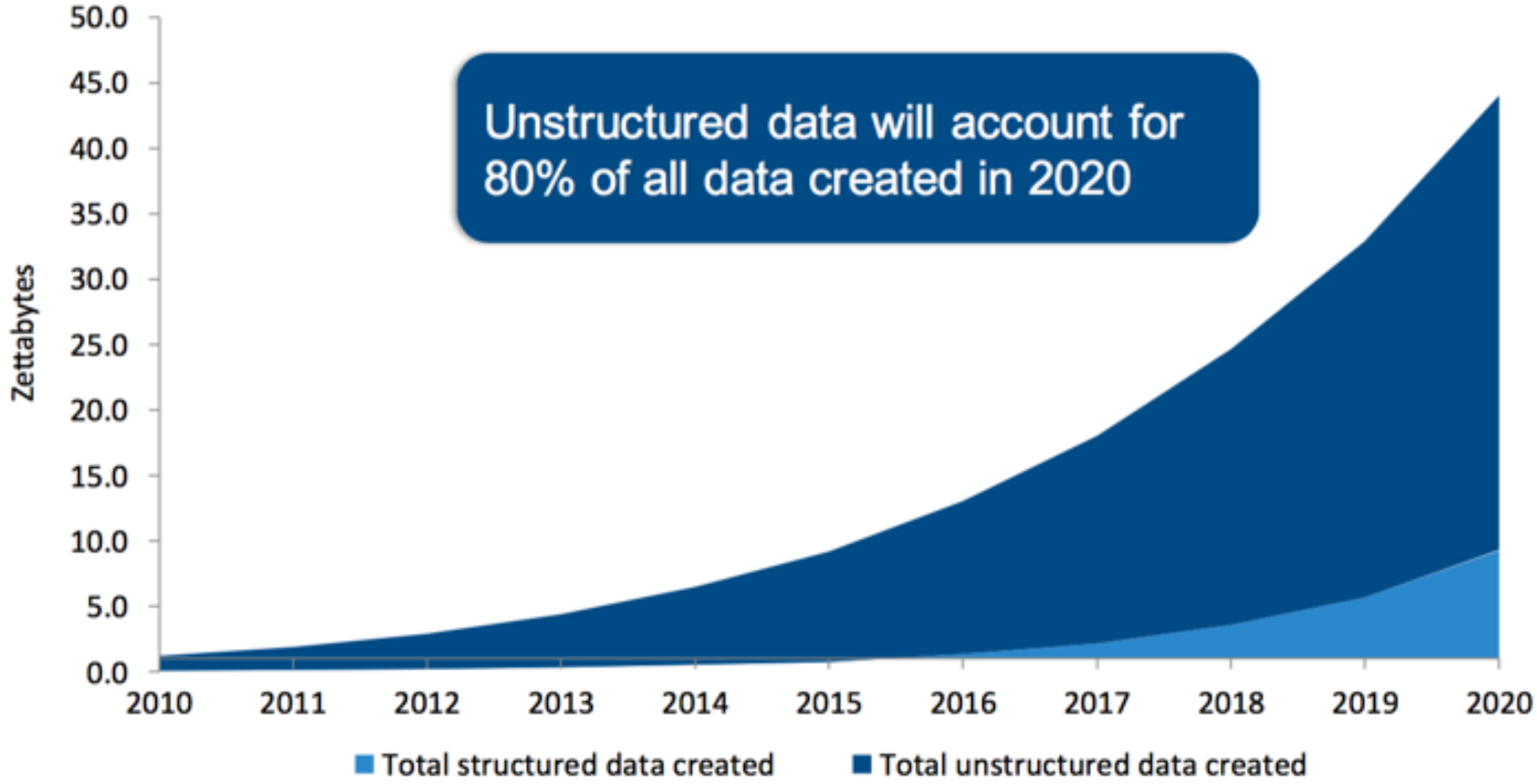
## Structured Data

- Well defined content
- Easily Understood
- Stored in RDBMS
- Easy to enter, store and analyze
- Example: data in database table (customer data, sales data, sensor data)

## Unstructured Data

- Structure not obvious
- Process data to understand
- RDBMS not a good fit
- Difficult and costly to analyze
- Example: email, videos, audio, web pages, social media feeds, presentation

## Capacity Growth by Data Type



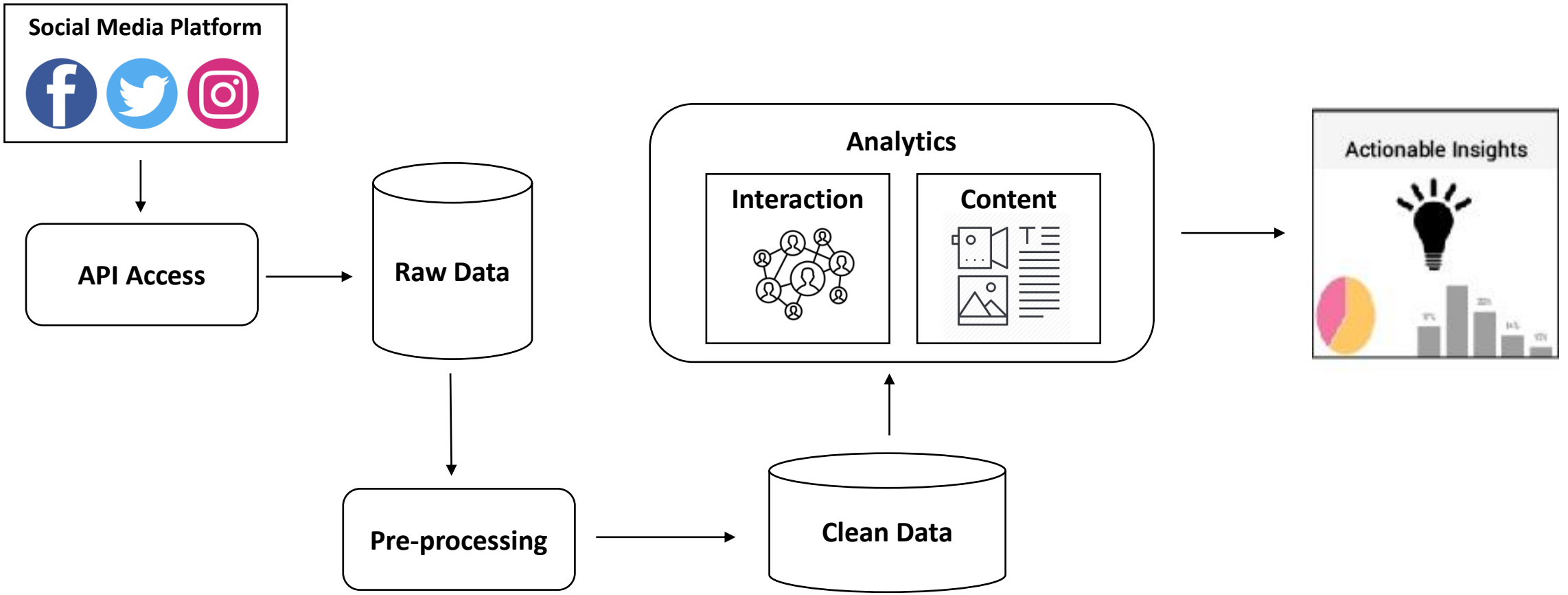

Source: IDC, 2016

https://www.ibm.com/blogs/cloud-computing/2016/12/13/idc-stacks-top-object-storage-vendors/

# Learning Outline

- Social Media Analytic Definition
- Background
- The Benefits of Social Media Analytic
- Challenges and Opportunities
- Understanding Data
- **Social Media Analytic Workflow**
- Social Media Analytic Implementation

# Social Media Analytics Workflow

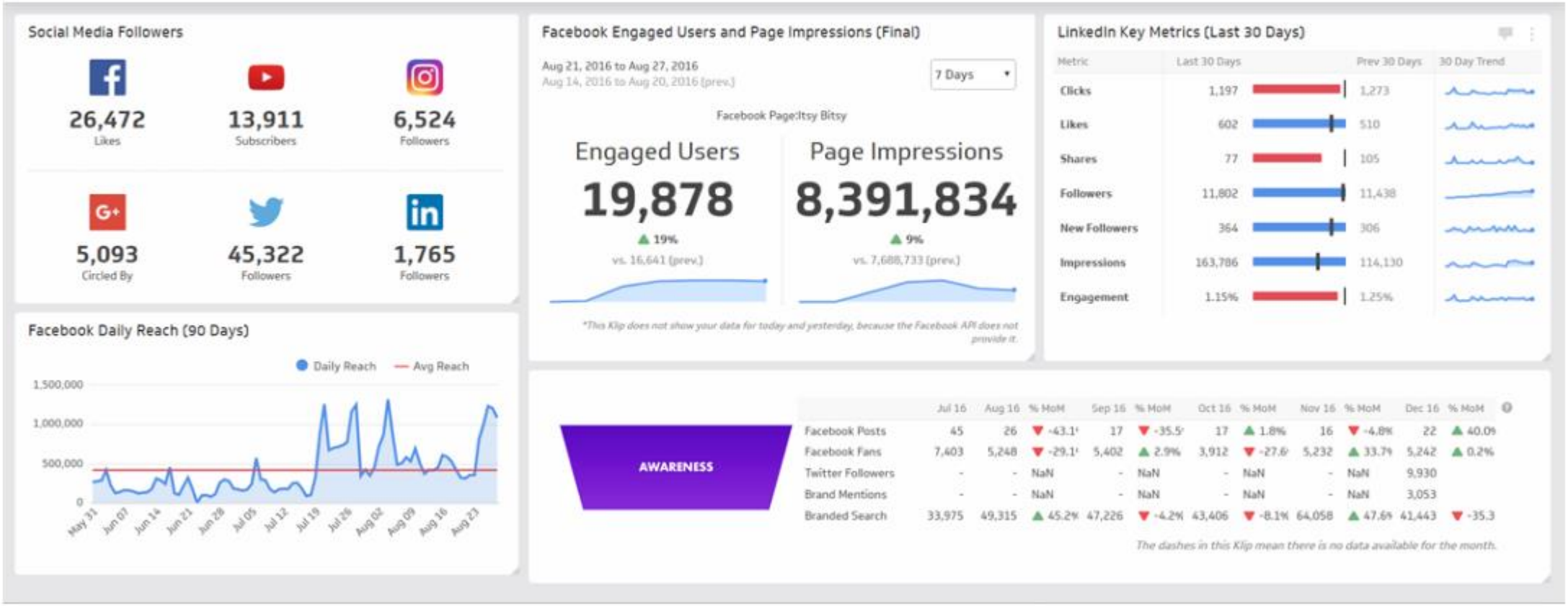


# Learning Outline

- Social Media Analytic Definition
- Background
- The Benefits of Social Media Analytic
- Challenges and Opportunities
- Understanding Data
- Social Media Analytic Workflow
- **Social Media Analytic Implementation**

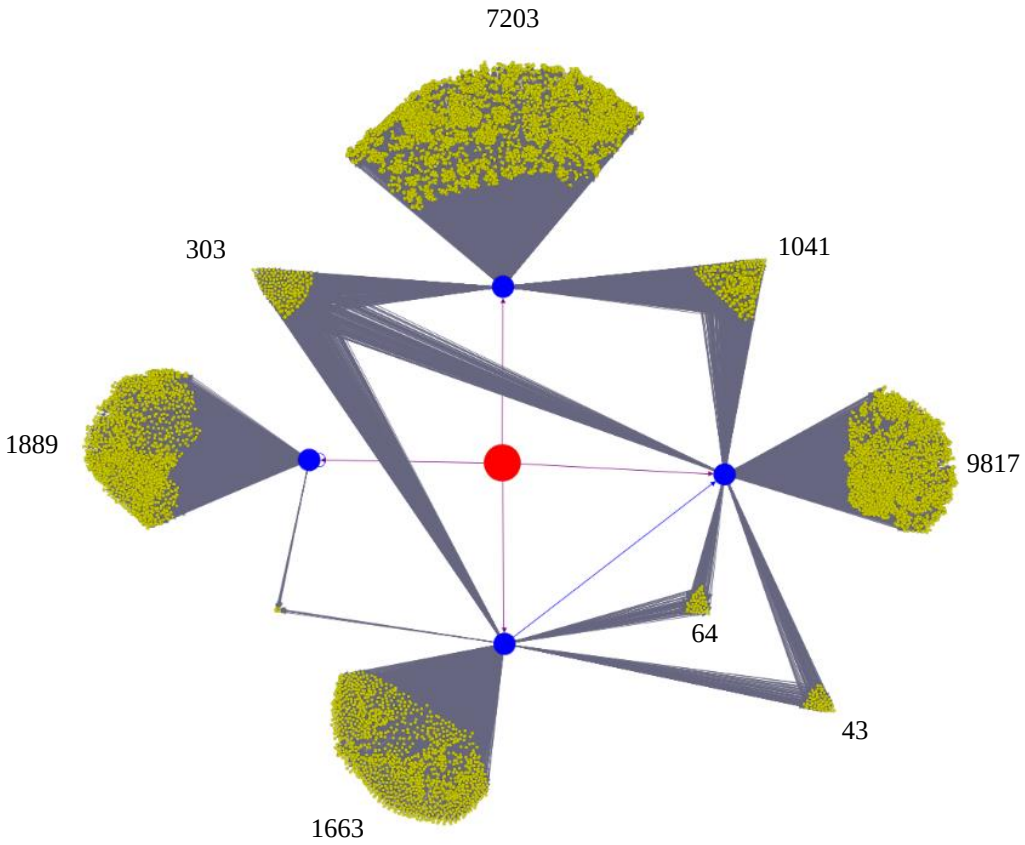
# Social Media Analytics Implementation

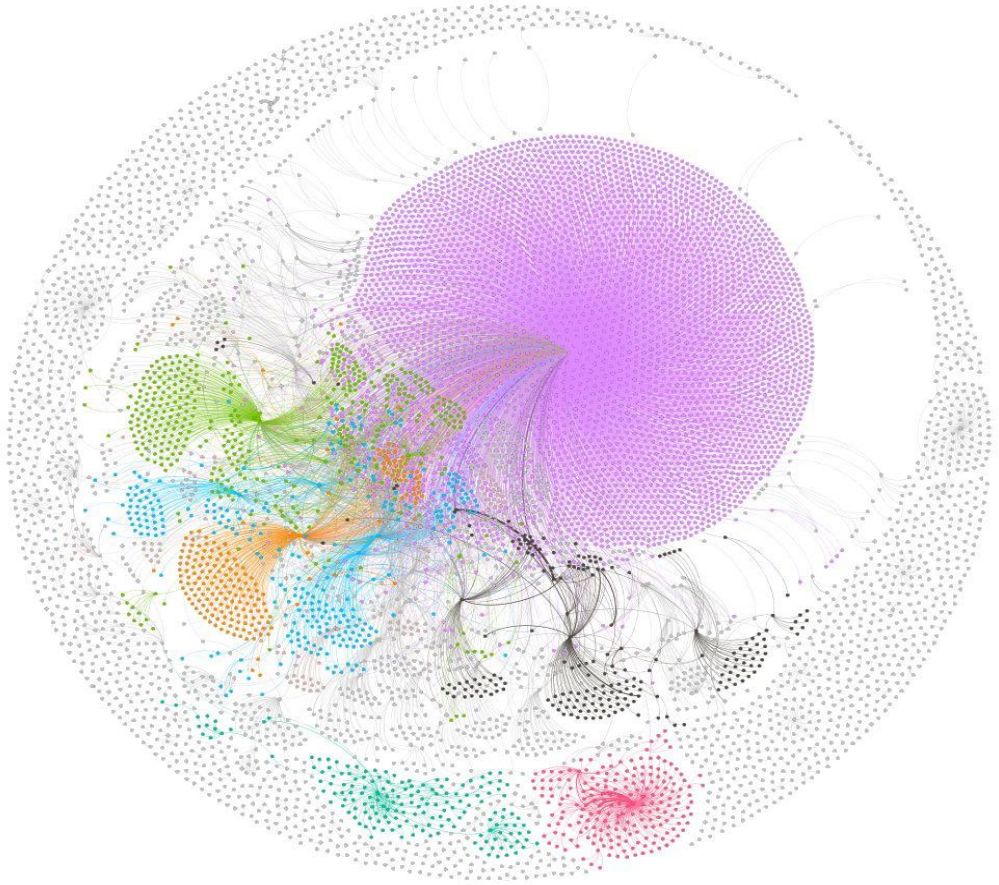
## Social Media Metrics



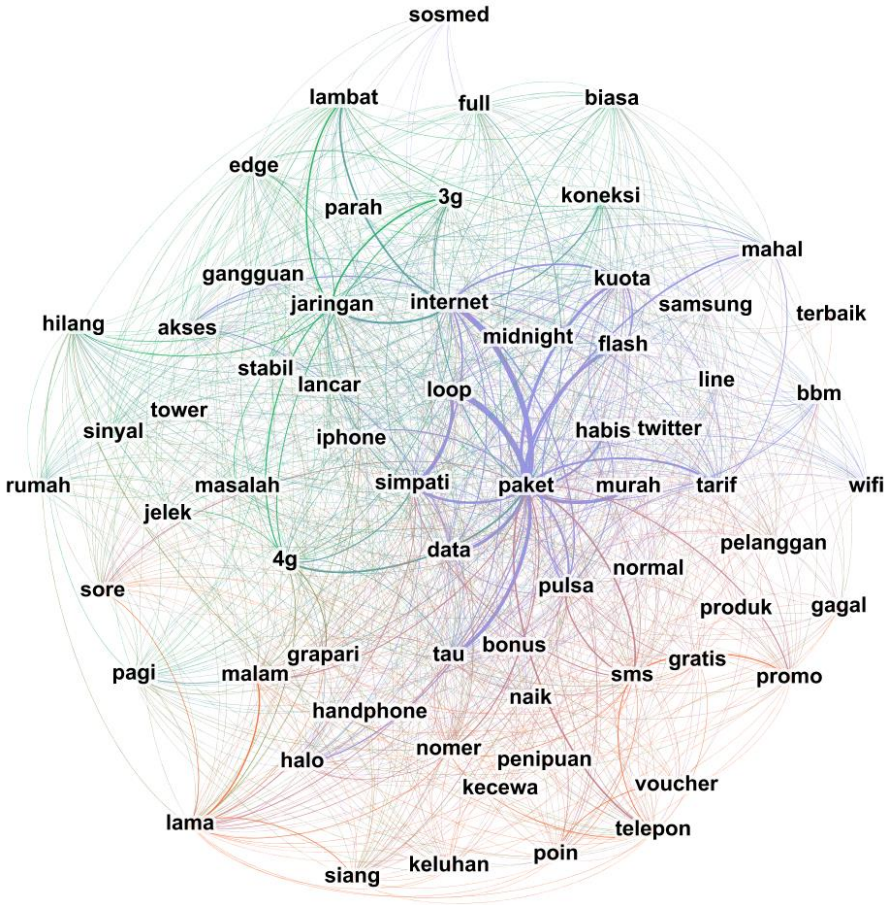

Telkom University
DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

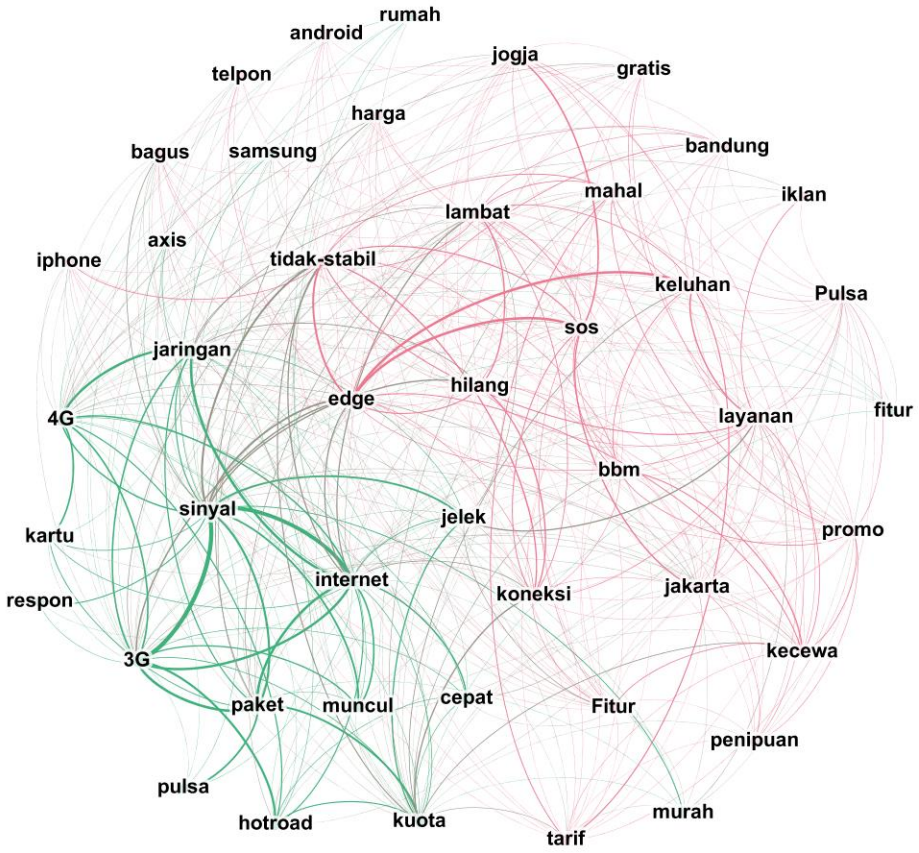
# Social Media Analytics Implementation



# Social Media Analytics Implementation

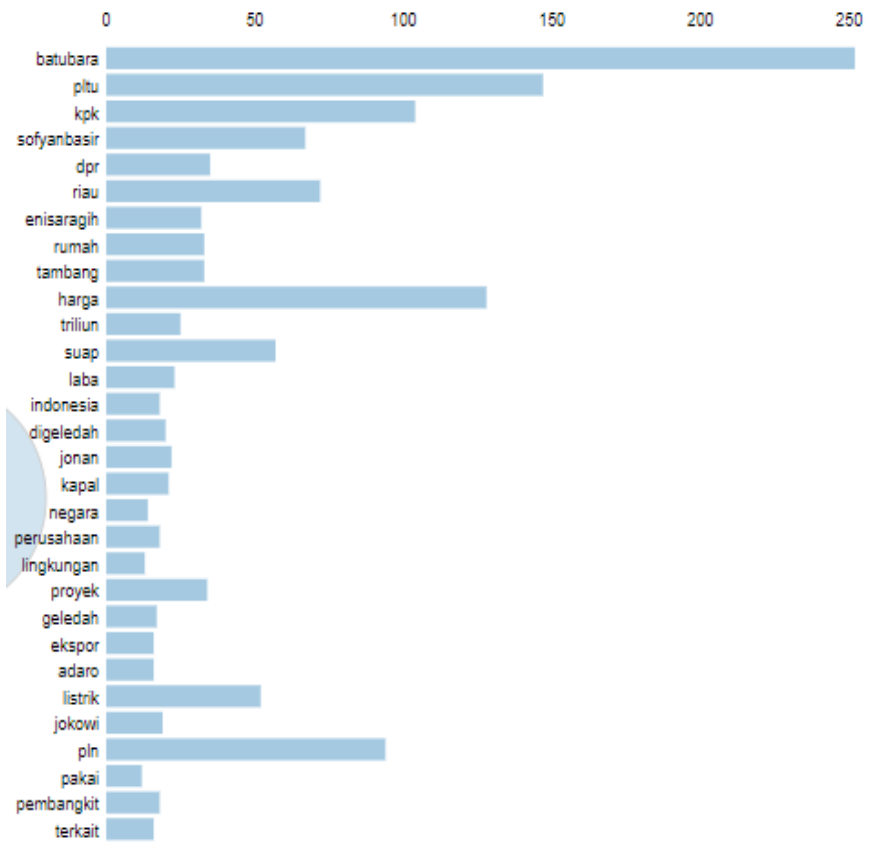


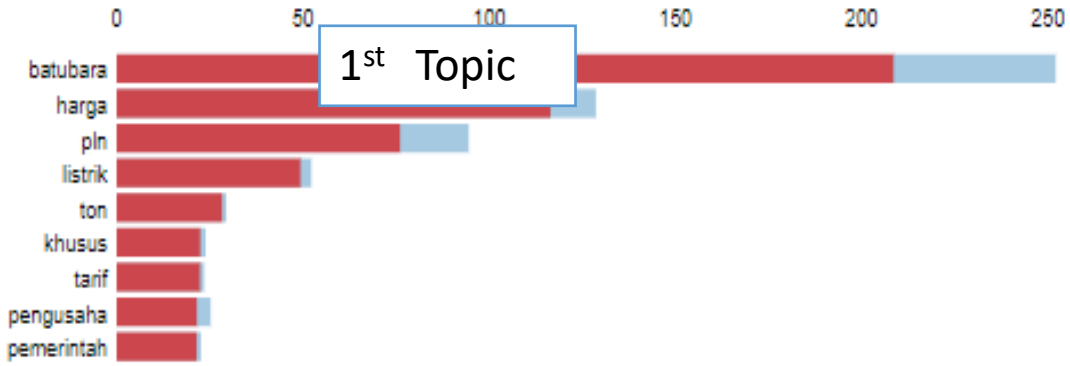




Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Social Media Analytics Implementation



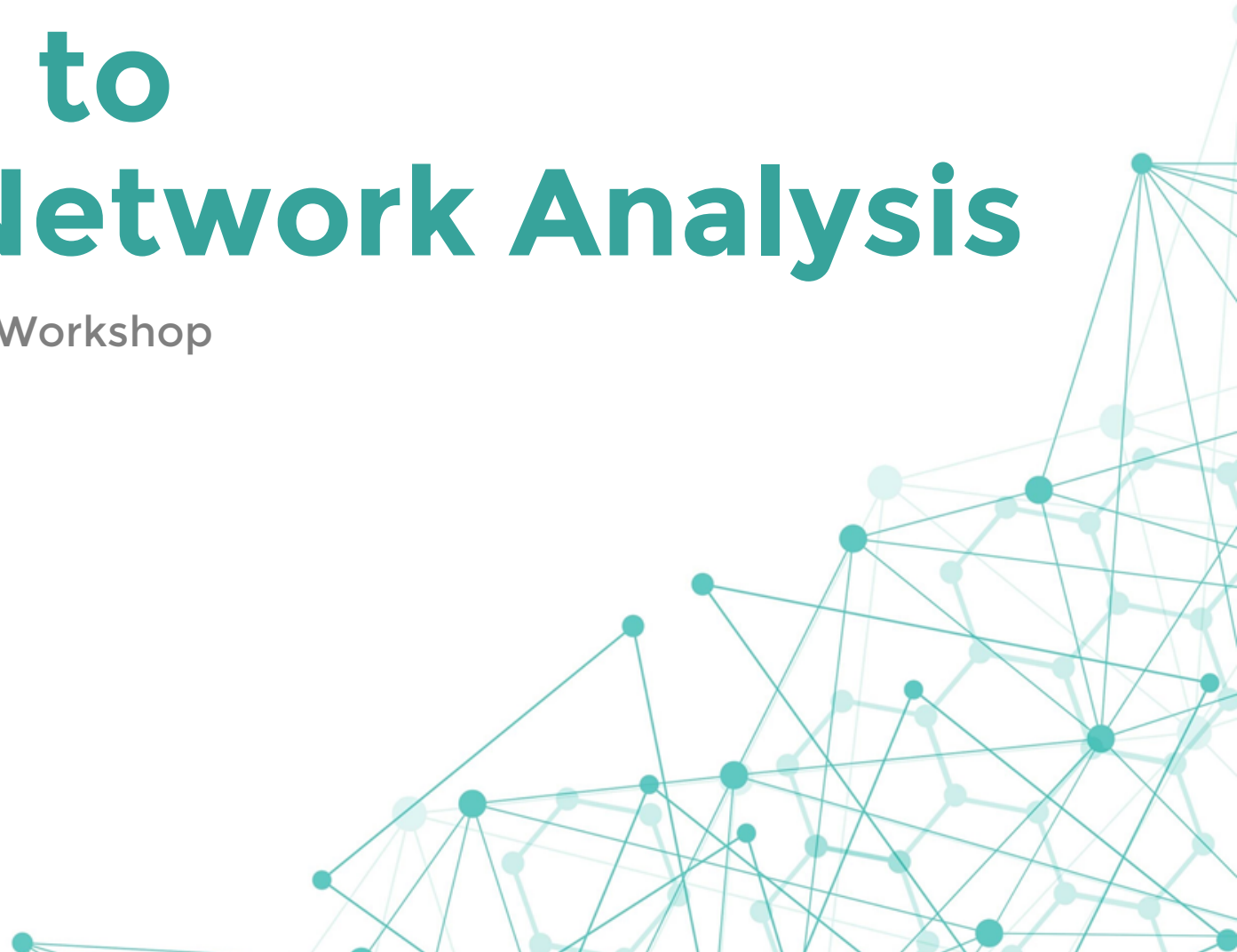

Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# 2

# Say Hai to Social Network Analysis

Social Media Analytic Workshop

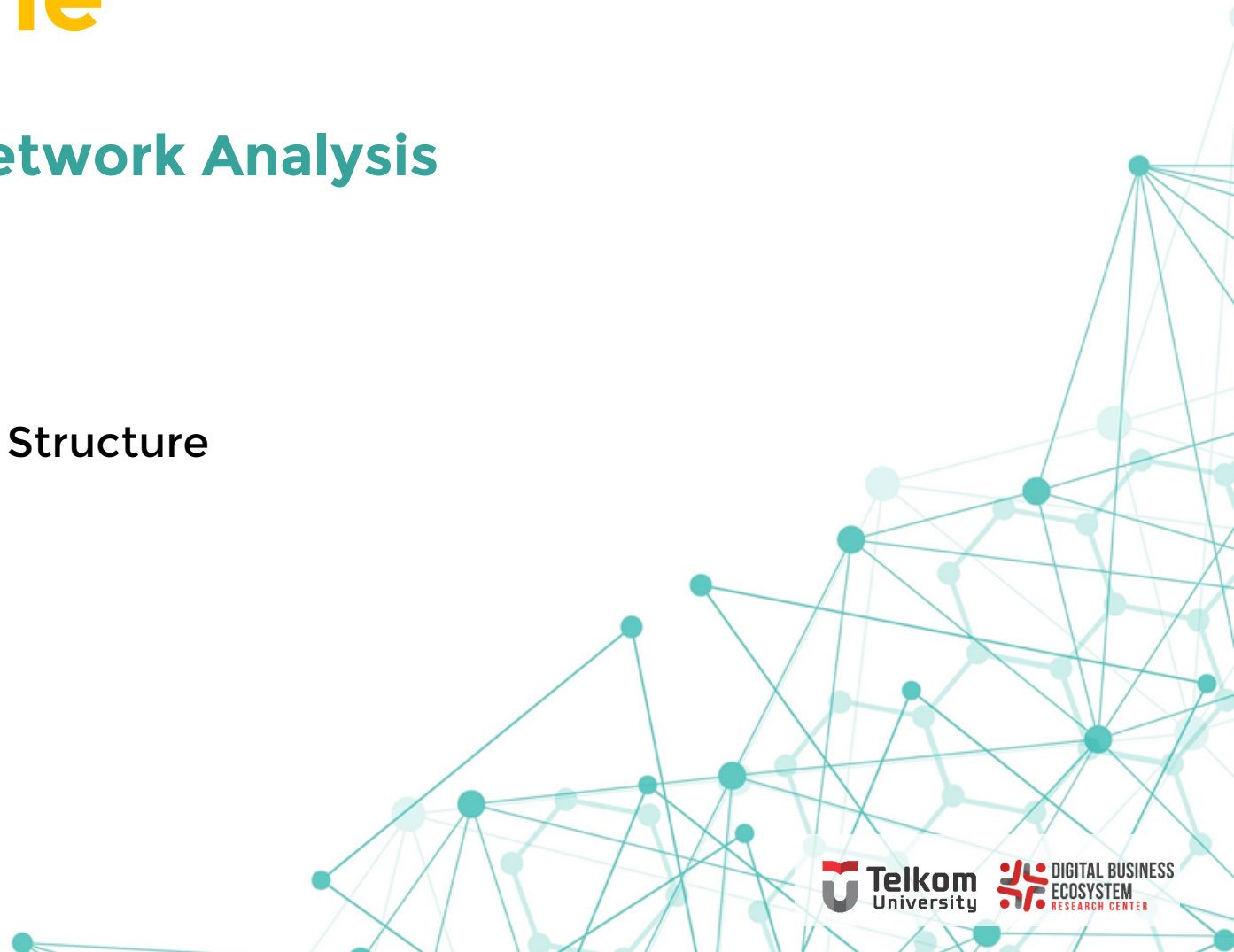


Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Learning Outline

- **Handshake with Social Network Analysis**
- Network Representation
- Tie Strength Identification
- Influencer Identification
- Measuring Overall Social Network Structure



# Network

- Powerful for describing complex systems by explaining interconnection between elements.
- Can be adopted universally in various fields e.g. mathematics, computer science, economics, sociology, chemistry, biology, etc.
- It has become essential to understand the real-world systems and gaining complex interactions knowledge

# Social Network Analysis

- Build upon:
  - Nodes as users
  - Edges as interaction flow between users
  - Graph type to indicate the nature of interaction
  - Weight to indicate the importance level of an interaction
- Focused on relationships and interconnected behaviors of various entities e.g. objects, people and organizations.
- The more specific area associated with SNA is the dynamic network behavior study formally known as Dynamic Network Analysis (DNA).

# Learning Outline

- Handshake with Social Network Analysis
- **Network Representation**
- Tie Strength Identification
- Influencer Identification
- Measuring Overall Social Network Structure

Telkom University
DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Network Representation (1)





**Can we study their interactions as a network ?**

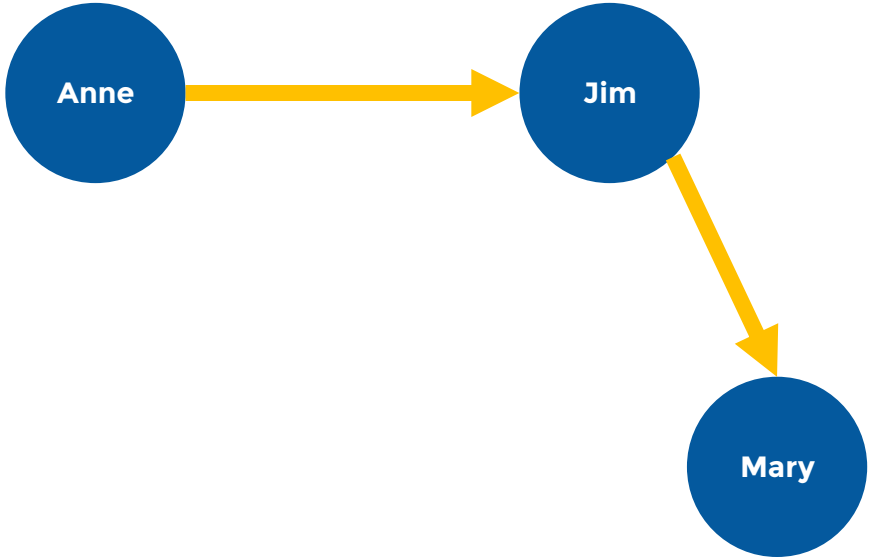
# Network Representation (2)



## Communication

Anne : Jim, tell Mary and John they're invited

# Network Representation (3)

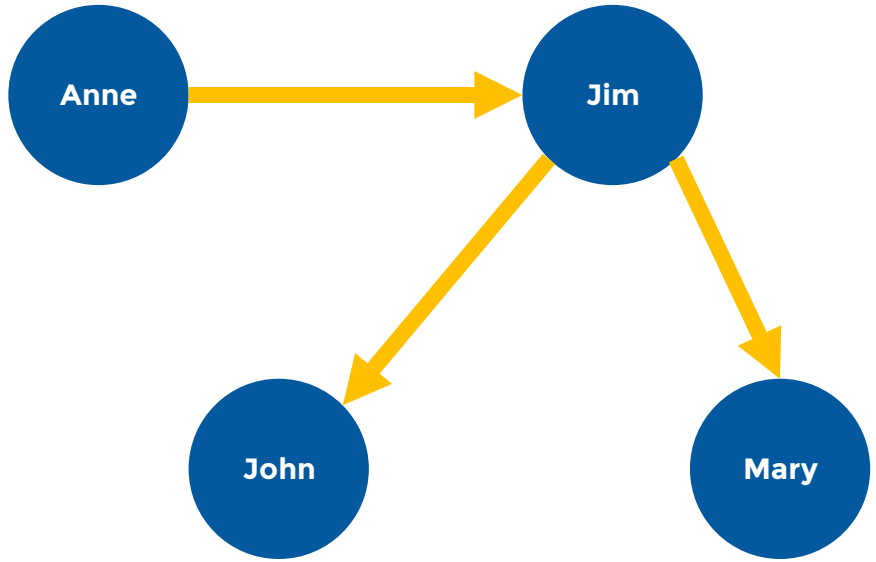




## Communication

Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

# Network Representation (4)

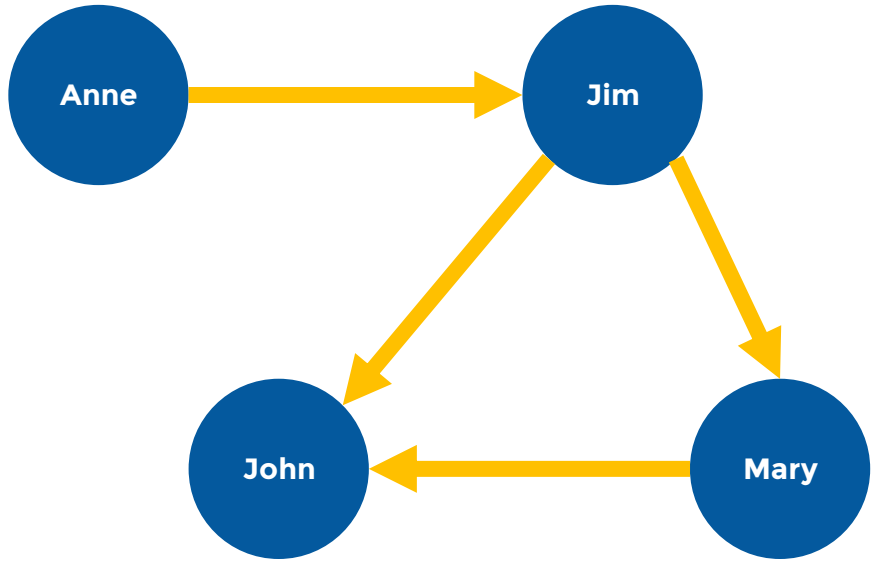

## Communication

Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

Jim : Mr. John, you should both come for dinner

# Network Representation (5)


## Communication

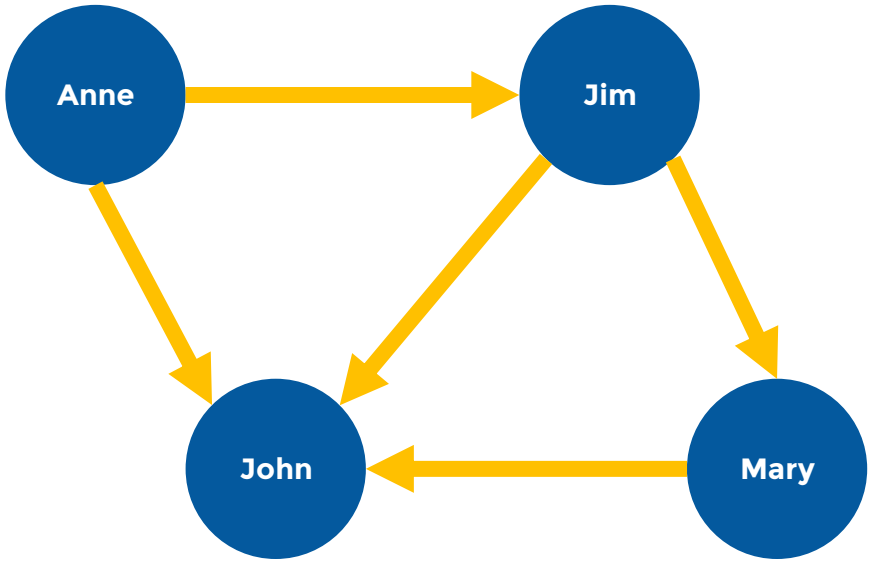
Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

Jim : Mr. John, you should both come for dinner

Mary : Dad, we are invited for tonight

# Network Representation (6)


## Communication

Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

Jim : Mr. John, you should both come for dinner

Mary : Dad, we are invited for tonight

Anne : John, did Jim tell you about the dinner? You must come

# Network Representation (7)


## Communication

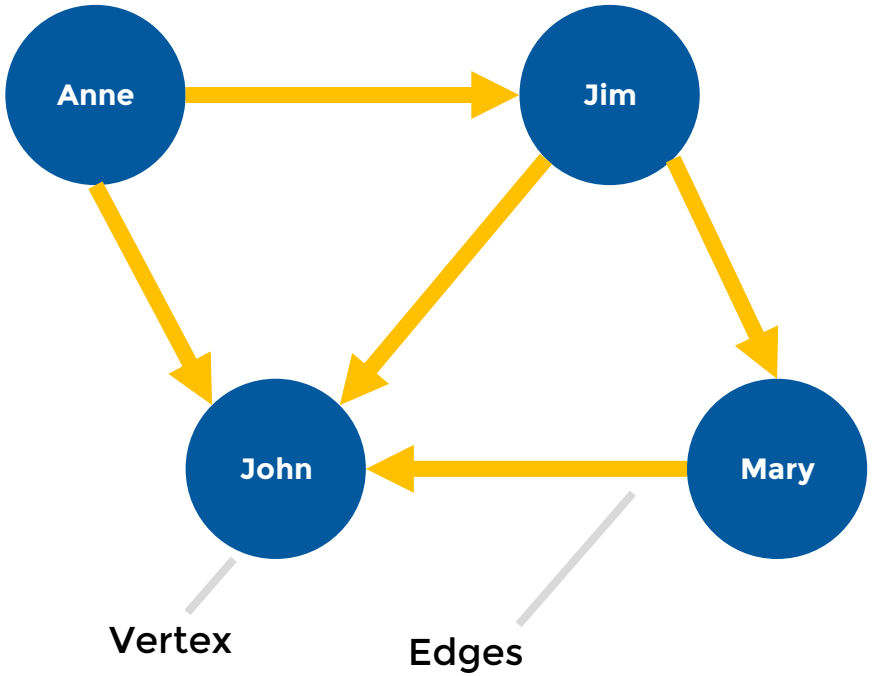
Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

Jim : Mr. John, you should both come for dinner

Mary : Dad, we are invited for tonight

Anne : John, did Jim tell you about the dinner? You must come



# Network Representation (8)

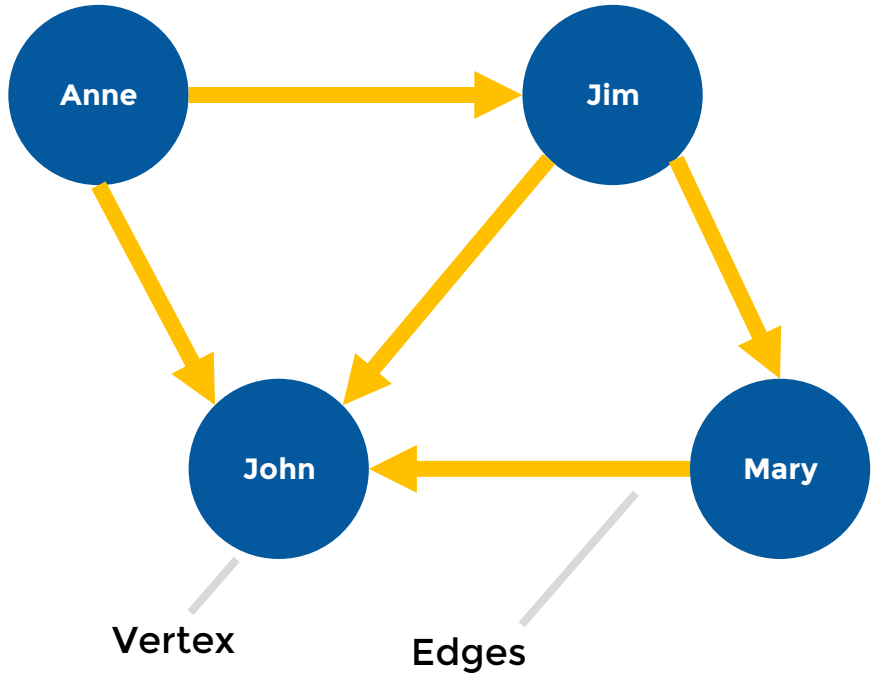
## Directed Network

***Edge List***

| Vertex | Vertex |
|---|---|
| Anne | Jim |
| Anne | John |
| Jim | John |
| Jim | Mary |
| Mary | John |

***Adjacency Matrix***

| Vertex | Anne | Jim | John | Mary |
|---|---|---|---|---|
| Anne | - | 1 | 1 | 0 |
| Jim | 0 | - | 1 | 1 |
| John | 0 | 0 | - | 0 |
| Mary | 0 | 0 | 1 | - |



# Network Representation (9)

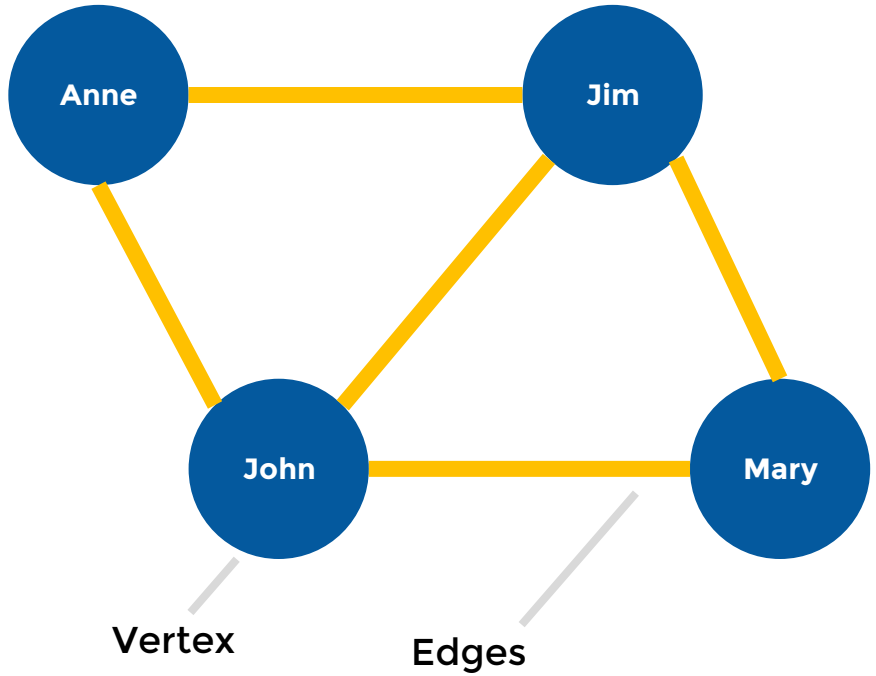
## Undirected Network

***Edge List***

| Vertex | Vertex |
|---|---|
| Anne | Jim |
| Anne | John |
| Jim | John |
| Jim | Mary |
| Mary | John |

***Adjacency Matrix***

| Vertex | Anne | Jim | John | Mary |
|---|---|---|---|---|
| Anne | - | 1 | 1 | 0 |
| Jim | 1 | - | 1 | 1 |
| John | 1 | 1 | - | 1 |
| Mary | 0 | 1 | 1 | - |



# Network Representation (10)

## Weighted Network

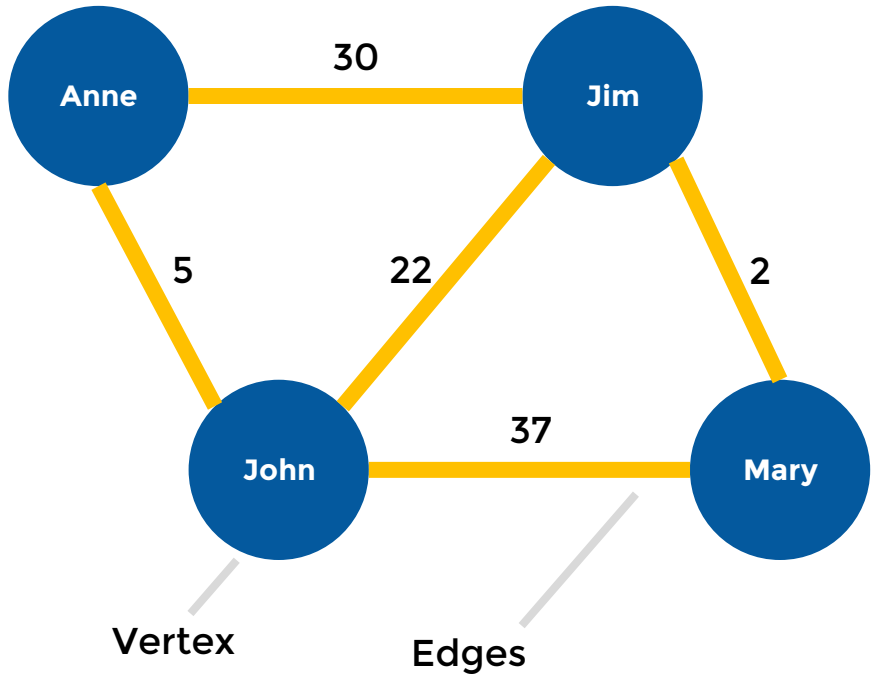
***Edge List***

| Vertex | Vertex | Weight |
|---|---|---|
| Anne | Jim | 30 |
| Anne | John | 5 |
| Jim | John | 22 |
| Jim | Mary | 2 |
| Mary | John | 27 |

***Adjacency Matrix***

| Vertex | Anne | Jim | John | Mary |
|---|---|---|---|---|
| Anne | - | 30 | 5 | 0 |
| Jim | 30 | - | 22 | 2 |
| John | 5 | 22 | - | 1 |
| Mary | 0 | 2 | 37 | - |

Weight could be
- Frequency of interactions in period of observation
- Number of items exchanged in period
- Individual perceptions of strength of relationship
- Cost of communications or exchange, e.g. distance



# Learning Outline

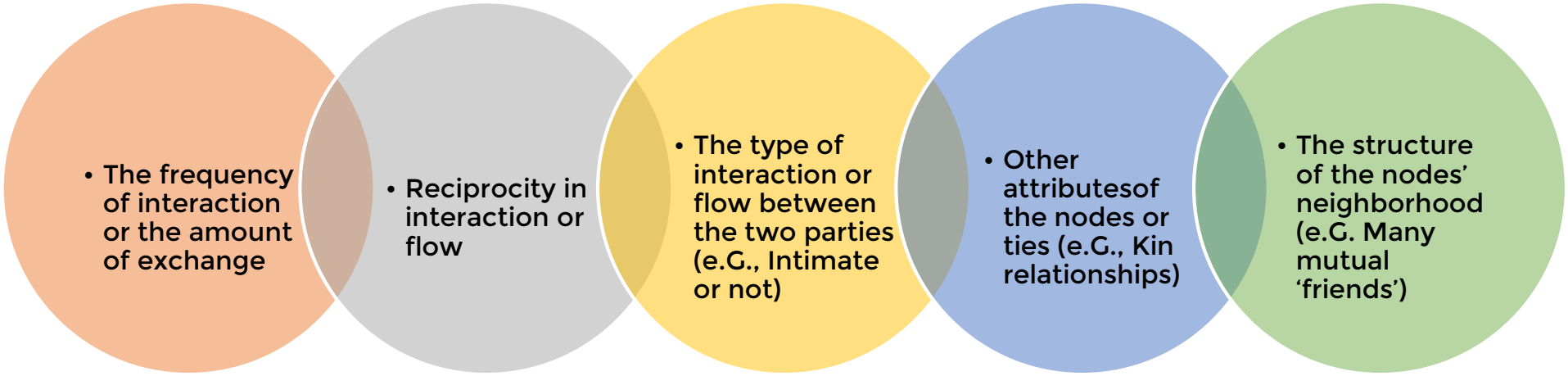
- Handshake with Social Network Analysis
- Network Representation
- **Tie Strength Identification**
- Influencer Identification
- Measuring Overall Social Network Structure

Telkom University
DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Edge Weight as Relationship Strength

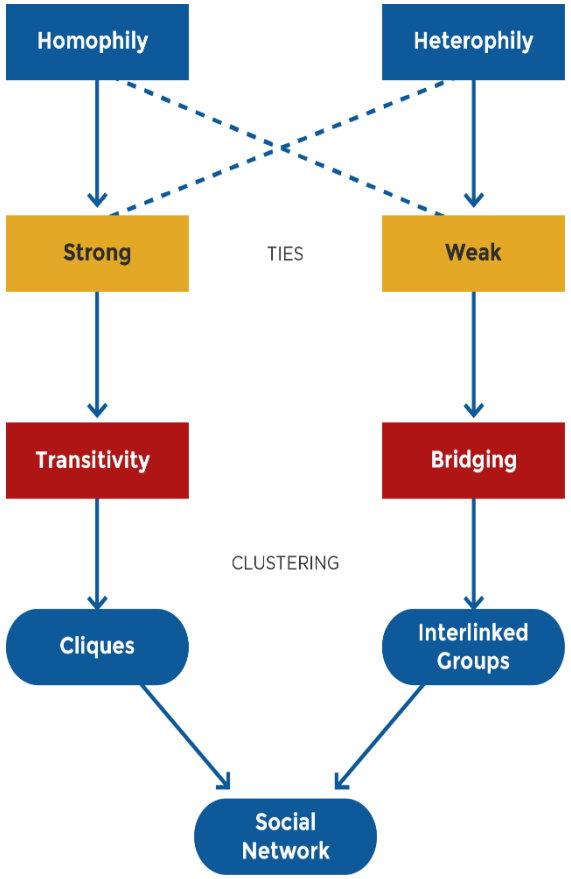
Edges can represent **interactions, flows of information or goods, similarities/affiliations,** or **social relations**

Specifically for social relations, a 'proxy' for the strength of a tie can be:




Surveys and interviews allows us to establish the existence of mutual or one-sided strength/affection with greater certainty, but proxies above are also useful

# Homophily, Transitivity, and Bridge



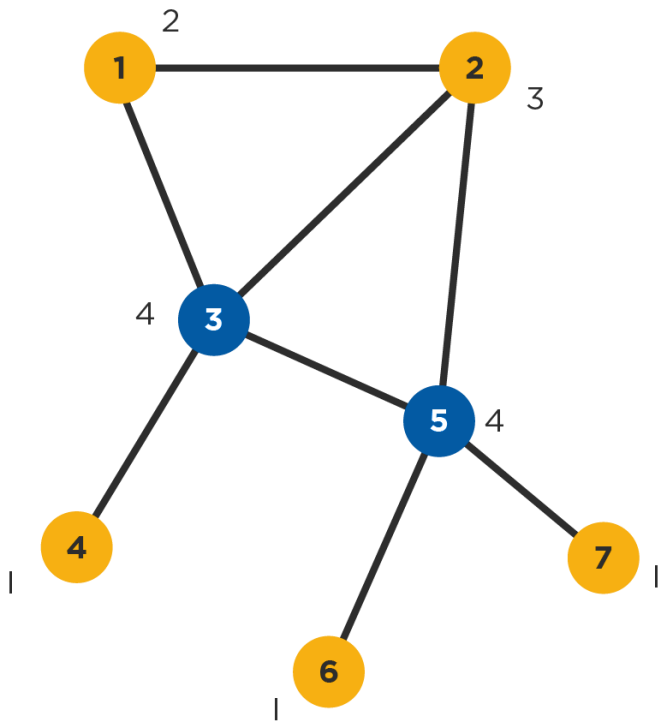

- Homophily is the tendency to relate to people with similar characteristics (status, beliefs, etc.)
- Transitivity in SNA is a property of ties: if there is a tie between A and B and one between B and C, then in a transitive network A and C will also be connected
- Bridges are nodes and edges that connect across groups

# Learning Outline

- Handshake with Social Network Analysis
- Network Representation
- Tie Strength Identification
- **Influencer Identification**
- Measuring Overall Social Network Structure

# Who is Important?

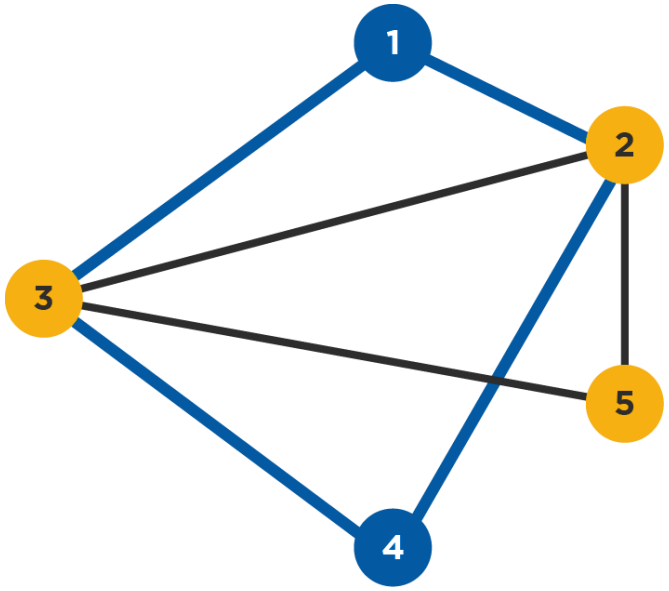
# Degree Centrality




Nodes 3 and 5 have the highest degree (4)

- A node's (in-) or (out-) degree is the number of links that lead into or out of the node
- In an undirected graph they are of course identical
- Often used as measure of a node's degree of connectedness and hence also influence and/or popularity
- Useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate 'neighborhood'

# Paths and Shortest Paths

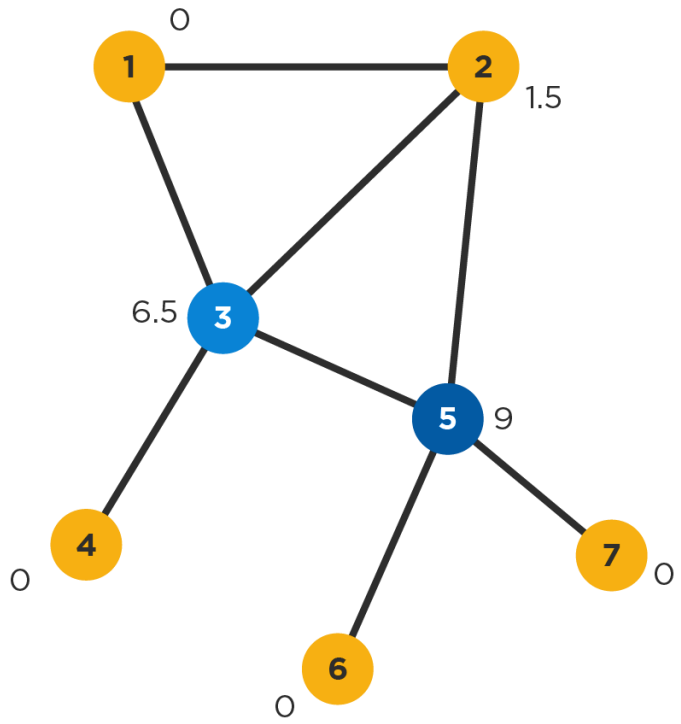



**Hypothetical Graph**

Shortest Path(s)

- A path between two nodes is any sequence of non-repeating nodes that connects the two nodes
- The shortest path between two nodes is the path that connects the two nodes with the shortest number of edges (also called the distance between the nodes)
- In the example to the right, between nodes 1 and 4 there are two shortest paths of length 2: {1,2,4} and {1,3,4}
- Other, longer paths between the two nodes are {1,2,3,4}, {1,3,2,4}, {1,2,5,3,4} and {1,3,5,2,4} (the longest paths)
- Shorter paths are desirable when speed of communication or exchange is desired (often the case in many studies, but sometimes not, e.g. in networks that spread disease)

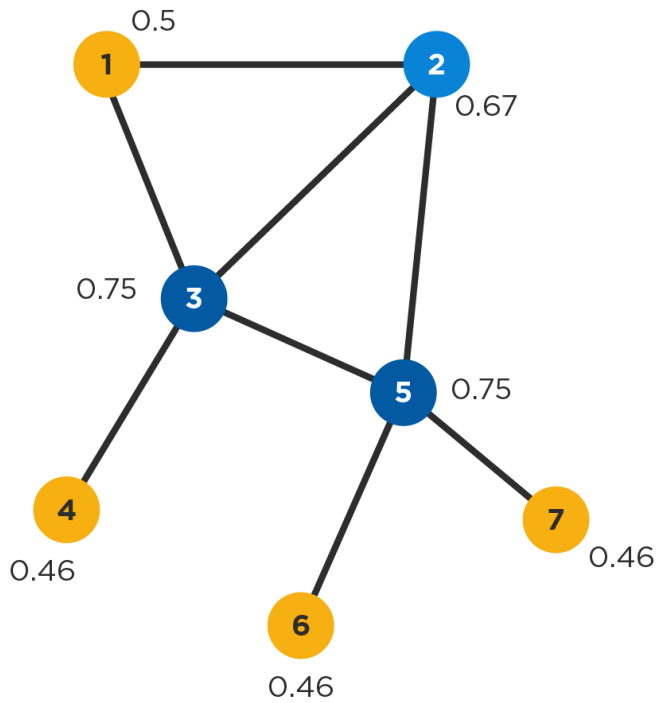
# Betweenness Centrality




Node 5 has higher betweenness centrality than 3

- For a given node v, calculate the number of shortest paths between nodes i and j that pass through v, and divide by all shortest paths between nodes i and j
- Sum the above values for all node pairs i,j
- Sometimes normalized such that the highest value is 1 or that the sum of all betweenness centralities in the network is 1
- Shows which nodes are more likely to be in communication paths between other nodes
- Also useful in determining points where the network would break apart (think who would be cut off if nodes 3 or 5 would disappear)

# Closeness Centrality

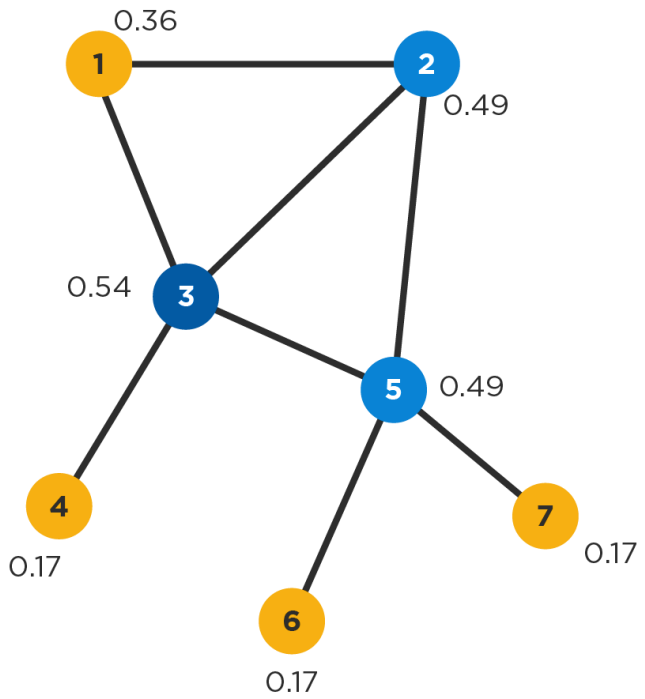



Note: Sometimes closeness is calculated without taking the reciprocal of the mean shortest path length. Then lower values are 'better'.

- Calculate the mean length of all shortest paths from a node to all other nodes in the network (i.e. how many hops on average it takes to reach every other node)
- Take the reciprocal of the above value so that higher values are 'better' (indicate higher closeness) like in other measures of centrality
- It is a measure of *reach*, i.e. the speed with which information can reach other nodes from a given starting node

Nodes 3 and 5 have the highest (i.e. best) closeness, while node 2 fares almost as well

# Eigenvector Centrality




Note: The term 'eigenvector' comes from mathematics (matrix algebra), but it is not necessary for understanding how to interpret this measure

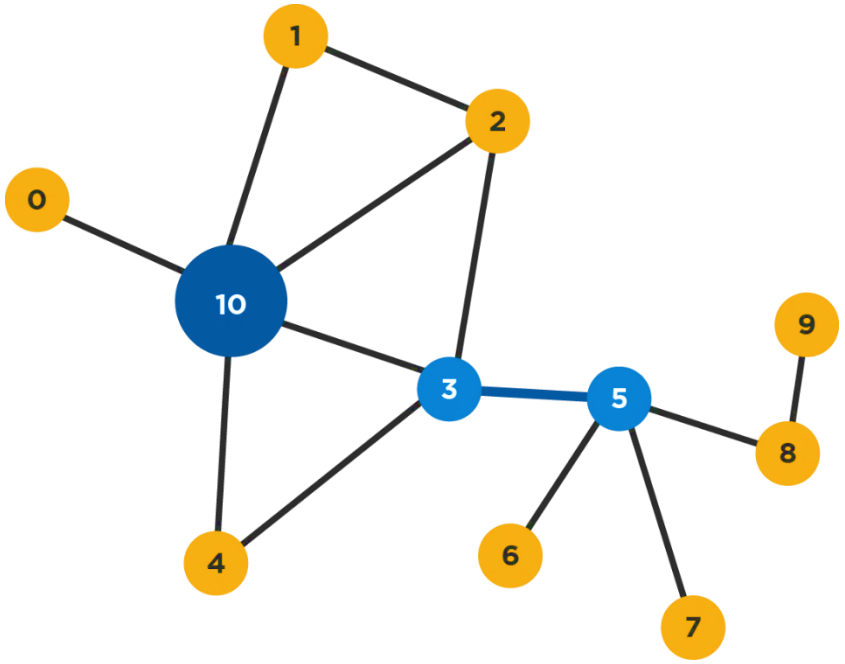
- A node's eigenvector centrality is proportional to the sum of the eigenvector centralities of all nodes directly connected to it
- In other words, a node with a high eigenvector centrality is connected to other nodes with high eigenvector centrality
- This is similar to how Google ranks web pages: links from highly linked-to pages count more
- Useful in determining who is connected to the most connected nodes

Node 3 has the highest eigenvector centrality, closely followed by 2 and 5

# Interpretation of Measures

| Centrality measure | Interpretation in social networks |
| --- | --- |
| Degree | How many people can this person reach directly? |
| Betweenness | How likely is this person to be the most direct route between two people in the network? |
| Closeness | How fast can this person reach everyone in the network? |
| Eigenvector | How well is this person connected to other well-connected people? |

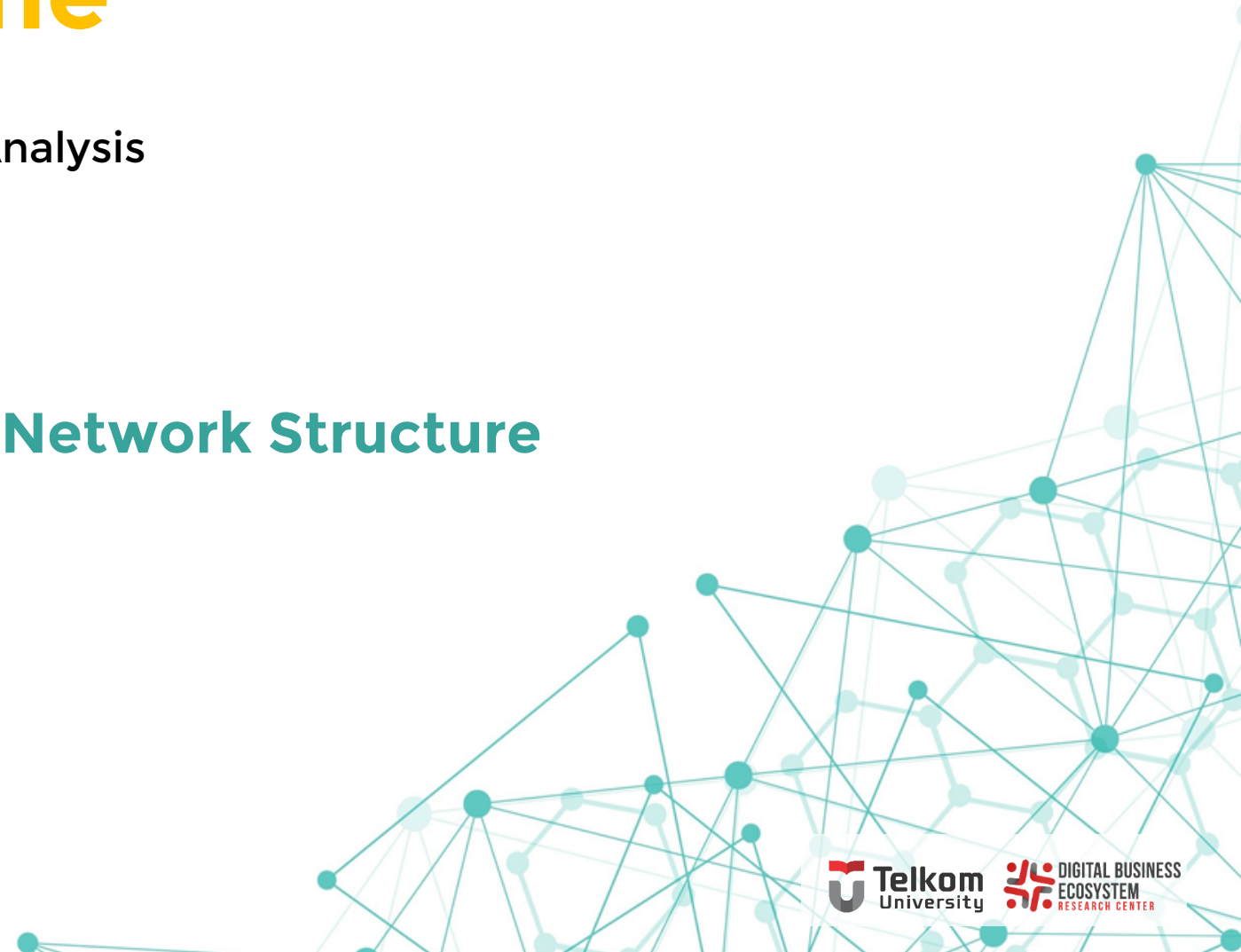
# Identifying Sets of Key Players




- In the network to the right, node 10 is the most central according to degree centrality
- But nodes 3 and 5 together will reach more nodes
- Moreover the tie between them is critical; if severed, the network will break into two isolated sub-networks
- It follows that other things being equal, players 3 and 5 together are more ‘key’ to this network than 10
- Thinking about sets of key players is helpful!

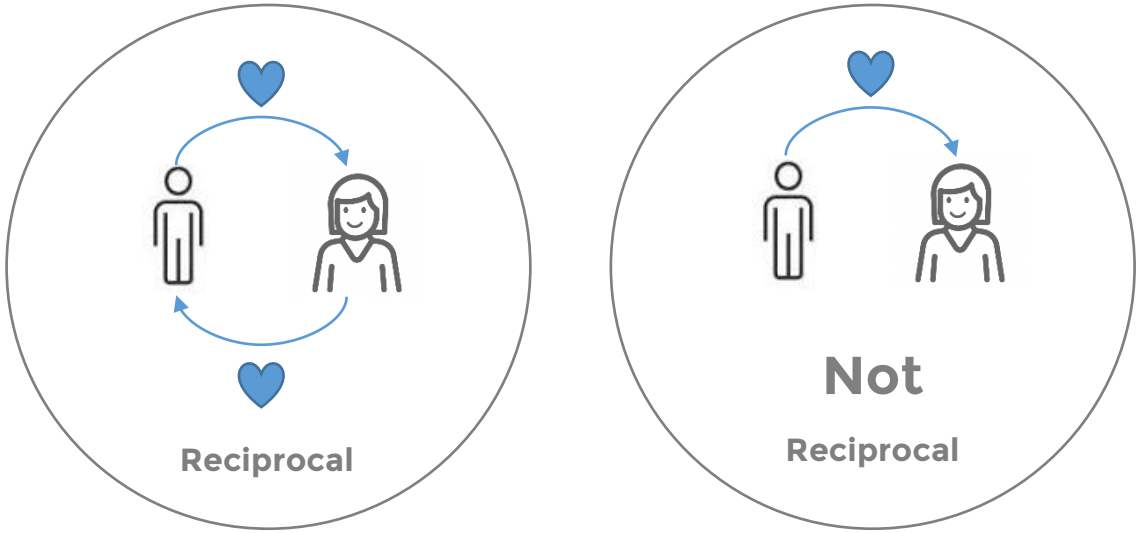
# Learning Outline

- Handshake with Social Network Analysis
- Network Representation
- Tie Strength Identification
- Influencer Identification
- **Measuring Overall Social Network Structure**



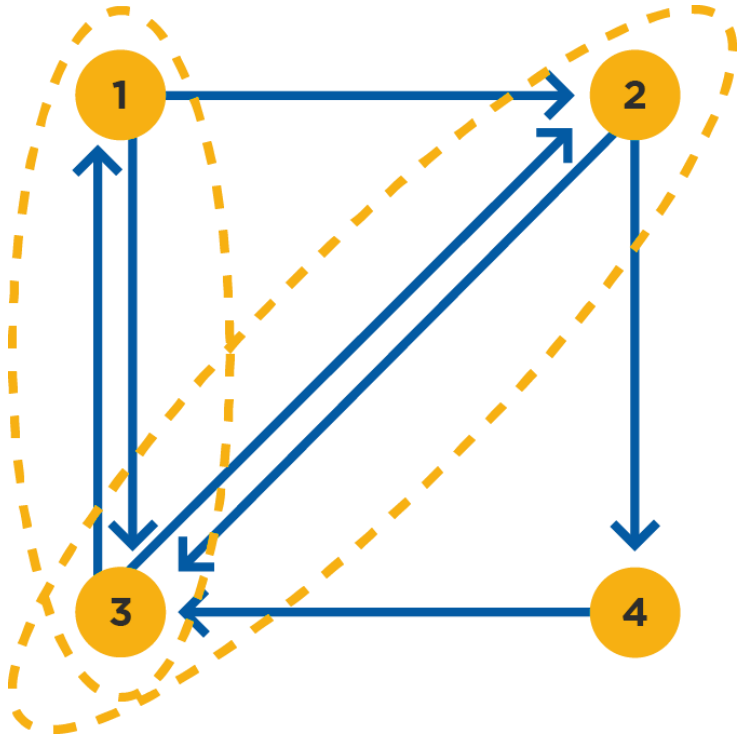
Telkom University
DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Reciprocity (degree of)




The ratio of the number of reciprocated relations over the total number of relations

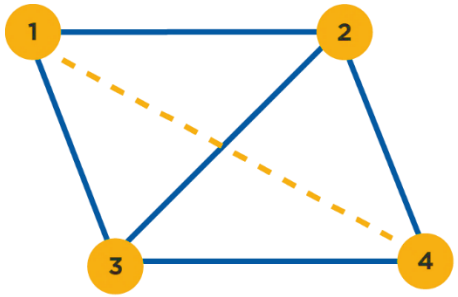
# Reciprocity (degree of)



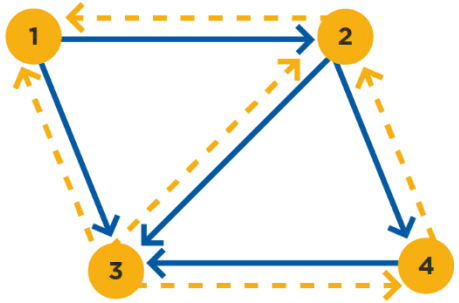

Reciprocity for network = 0.4

- The ratio of the number of relations which are reciprocated (i.e. there is an edge in both directions) over the total number of relations in the network
- …where two vertices are said to be related if there is at least one edge between them
- In the example to the right this would be 2/5=0.4 (whether this is considered high or low depends on the context)
- A useful indicator of the degree of mutuality and reciprocal exchange in a network, which relate to social cohesion
- Only makes sense in directed graphs

# Density




Density → $\frac{5}{6} = 0.83$



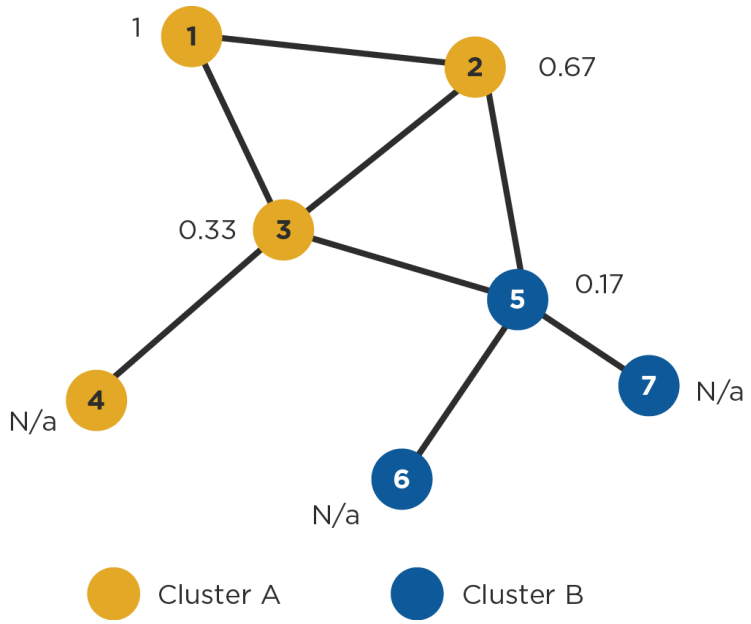

Density → $\frac{5}{12} = 0.42$

Edge present in network

Possible but not present

- A network's *density* is the ratio of the number of edges in the network over the total number of possible edges between all pairs of nodes (which is *n(n-1)/2*, where *n* is the number of vertices, for an undirected graph)
- In the example network to the right density=5/6=0.83 (i.e. it is a fairly *dense* network; opposite would be a *sparse* network)
- It is a common measure of how well connected a network is (in other words, how closely knit it is) – a perfectly connected network is called a *clique* and has density=1
- A directed graph will have half the density of its undirected equivalent, because there are twice as many possible edges, i.e. *n(n-1)*
- Density is useful in comparing networks against each other, or in doing the same for different regions within a single network

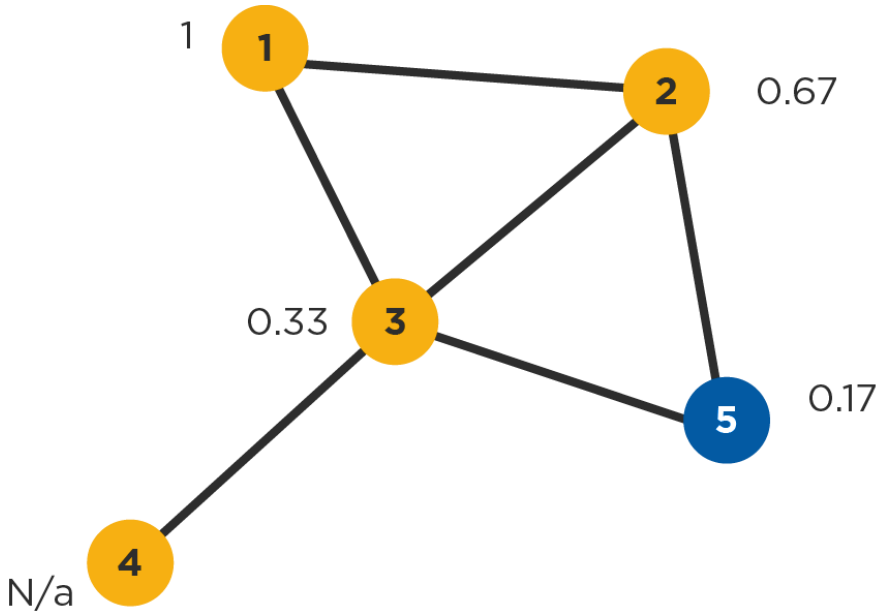
# Clustering



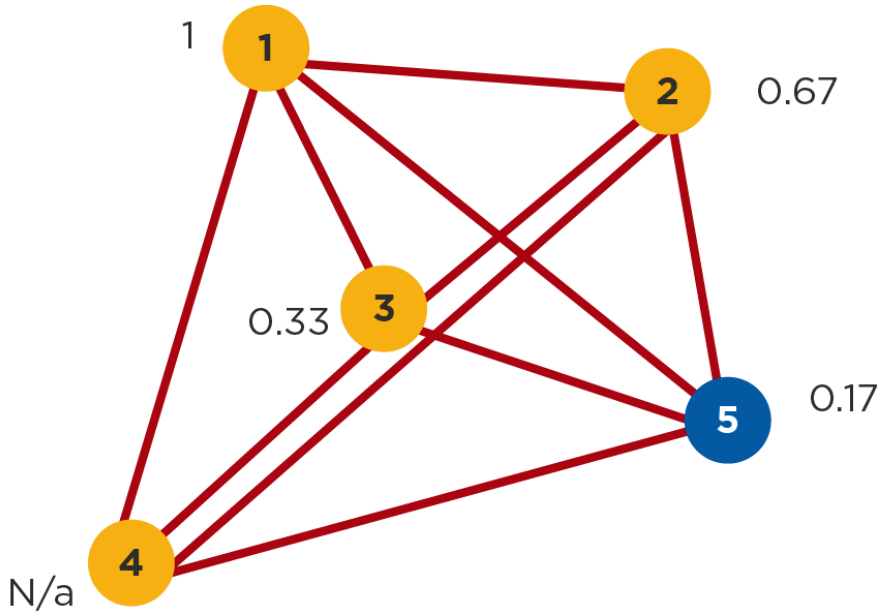

Network clustering coefficient = 0.375
(3 nodes in each triangle x 2 triangles = 6 closed triplets divided by 16 total)

- A node's *clustering coefficient* is the number of closed triplets in the node's neighborhood over the total number of triplets in the neighborhood. It is also known as *transitivity*.
- E.g., node 1 to the right has a value of 1 because it is only connected to 2 and 3, and these nodes are also connected to one another (i.e. the only triplet in the neighborhood of 1 is closed). We say that nodes 1,2, and 3 form a *clique*.
- Clustering algorithms identify clusters or 'communities' within networks based on network structure and specific clustering criteria (example shown to the right with two clusters is based on *edge betweenness*, an equivalent for edges of the betweenness centrality presented earlier for nodes)

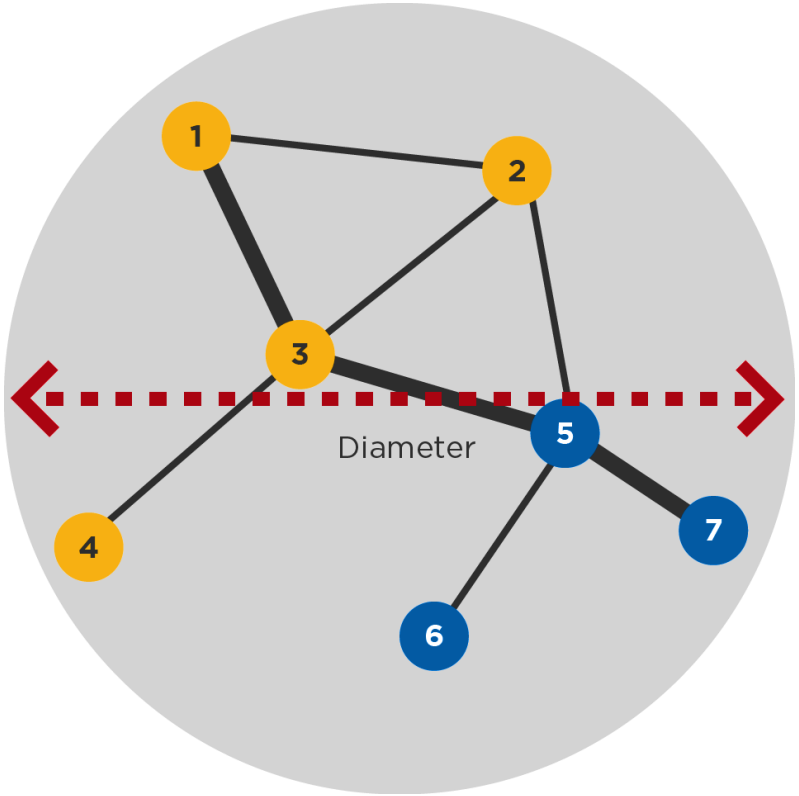
# Triplets of Node 3




## Possible Number of Triplets



# Average and Longest Distance



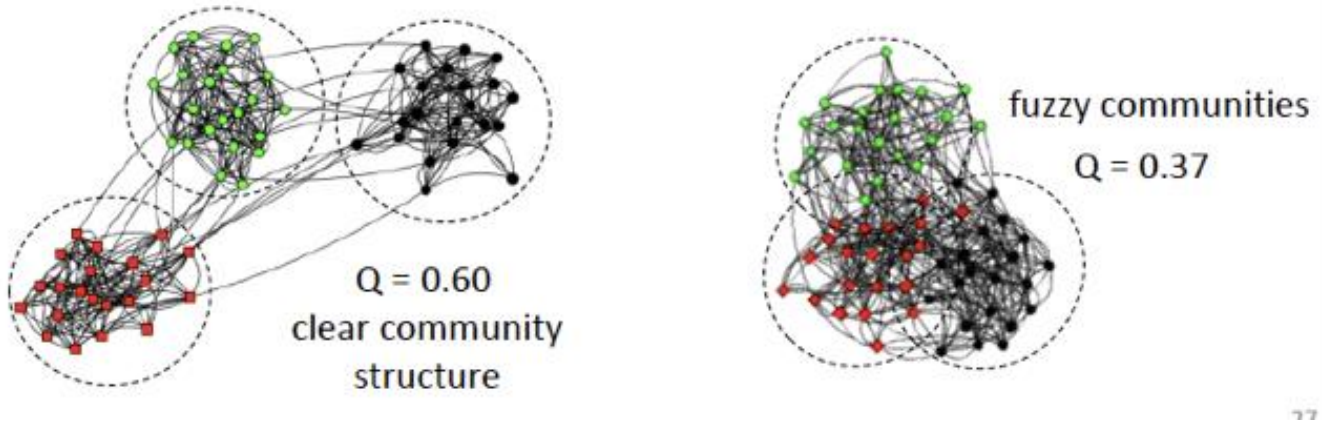

- The longest shortest path (distance) between any two nodes in a network is called the network's *diameter*
- The diameter of the network on the right is 3; it is a useful measure of the *reach* of the network (as opposed to looking only at the total number of vertices or edges)
- It also indicates how long it will take at most to reach any node in the network (sparser networks will generally have greater diameters)
- The average of all shortest paths in a network is also interesting because it indicates how far apart any two nodes will be on average (average *distance*)

# Modularity

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j)$$

Edges inside the community

Expected number of edges if i,j places at random

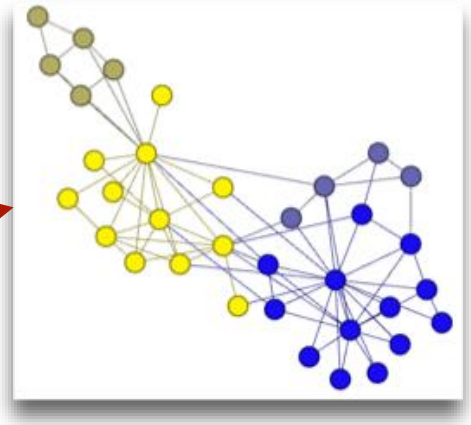


# Metric Simulation

*model network*

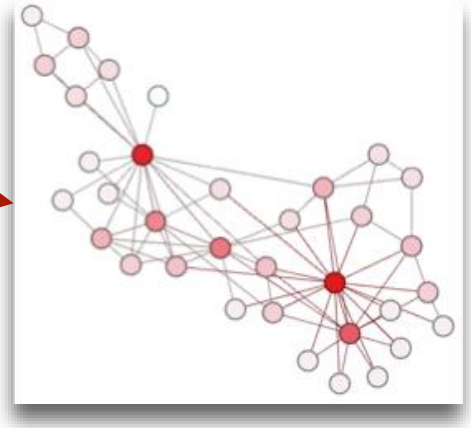
modularity

community detection



*community detection result*

centrality

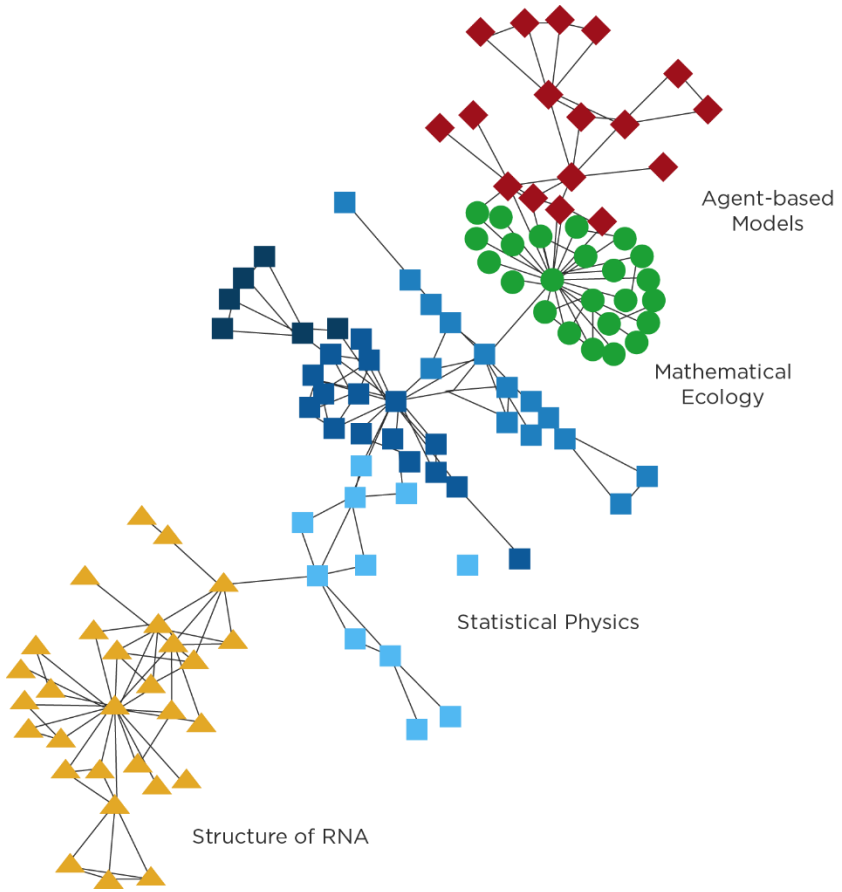


*degree centrality result*

Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Example : Finding Community




- Collaboration network of scientist at Santa Fe Institut (Girvan & Nirwan)
- 27| scientist (vertices) / 1 | 8 nodes from largest component edge = scientist coauthor one of more publications
- Komunitas : kumpulan titik titik dimana jumlah hubungan internal antar titik lebih besar dari pada jumlah hubungan dengan titik eksternal


## I just need my friend,

friend of my friend, friend of friend of my friend, friend of friend of friend of my friend, friend of friend of friend of friend of my friend, friend of friend of friend of friend of my friend, friend of friend of friend of friend of my friend,

## to contact him

This is Dian

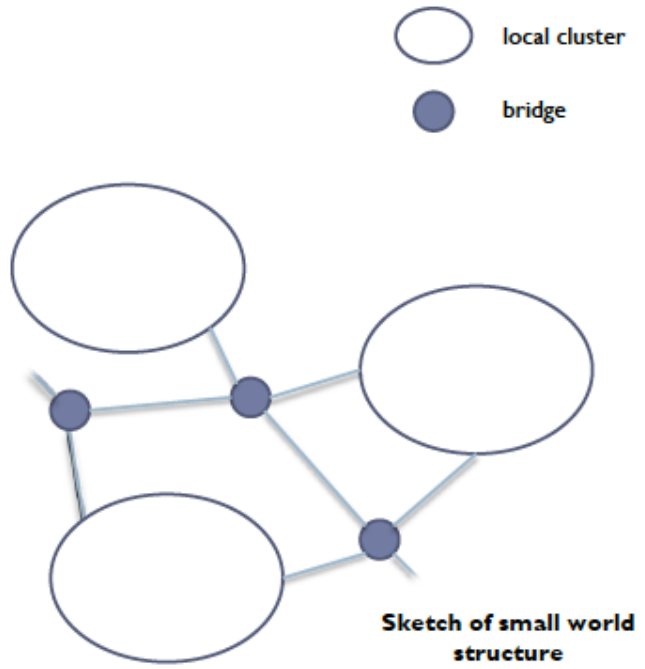

This is Peter Parker

Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Small Worlds

- A small world is a network that looks almost random but exhibits a significantly high clustering coefficient(nodes tend to cluster locally) and a relatively short average path length (nodes can be reached in a few steps
- It is a very common structure in social networks because of transitivity in strong social ties and the ability of weak ties to reach across clusters
- Such a network will have many clusters but also many bridges between clusters that help shorten the average distance between nodes



Sketch of small world structure

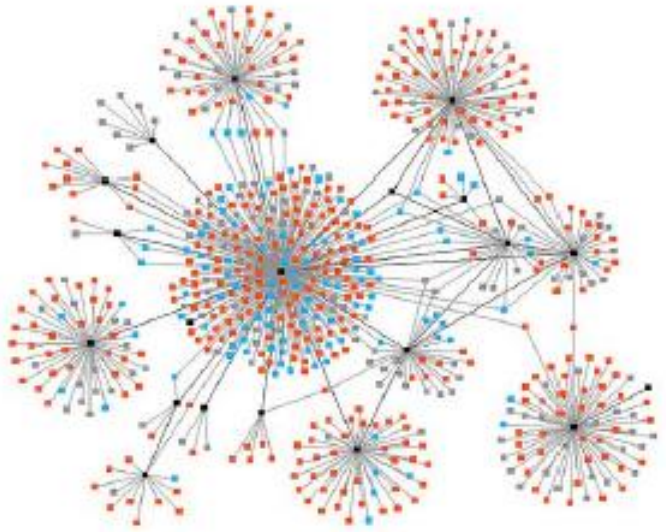
You may have heard of the famous “6 degrees” of separations

# Preferential Attachment

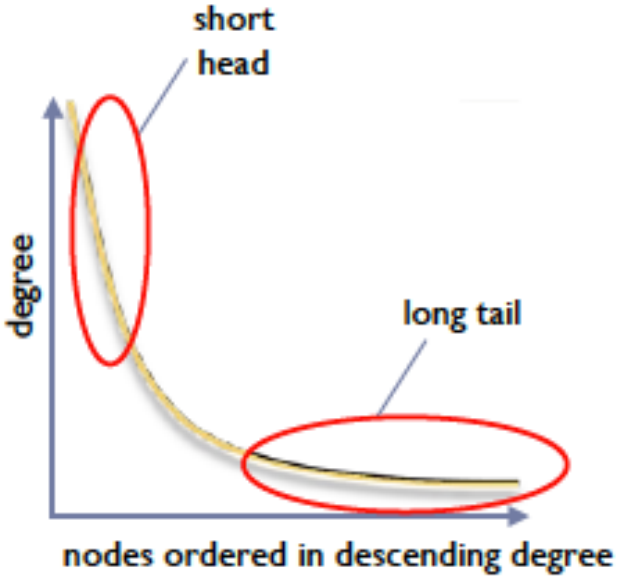
A property of some networks, where, during their evolution and growth in time, a the great majority of new edges are to nodes with an already high degree; the degree of these nodes thus increases disproportionately, compared to most other nodes in the network

- The result is a network with few very highly connected nodes and many nodes with a low degree
- Such networks are said to exhibit a long-tailed degree distribution
- And they tend to have a small-world structure!

*(it turns out, transitivity and strong/weak tie characteristics are not neccessary to explain small world structures, but they are common and can also lead to such structures)*



**Example of network with preferential attachment**



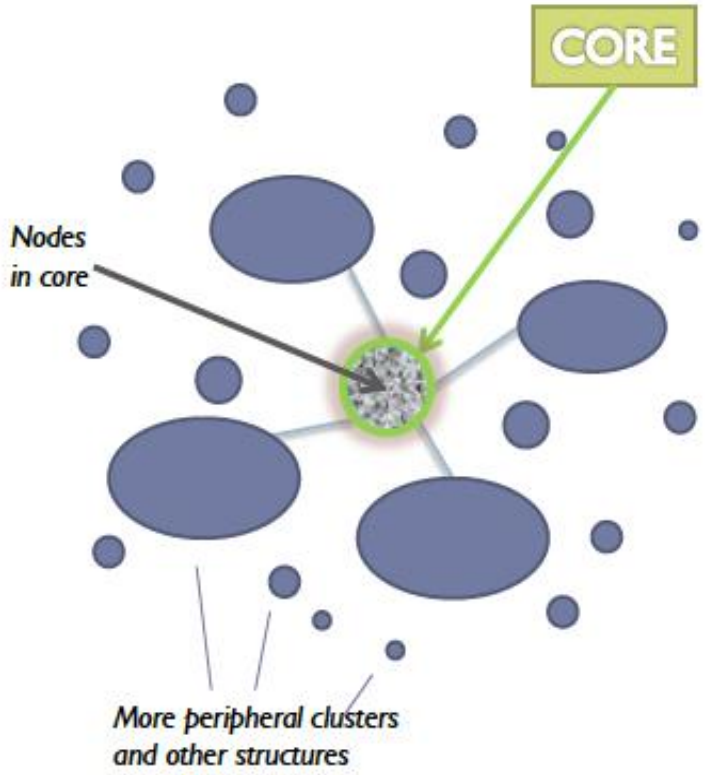

**Skectch of long-tailed degree distributions**

# Reasons for Preferential Attachment

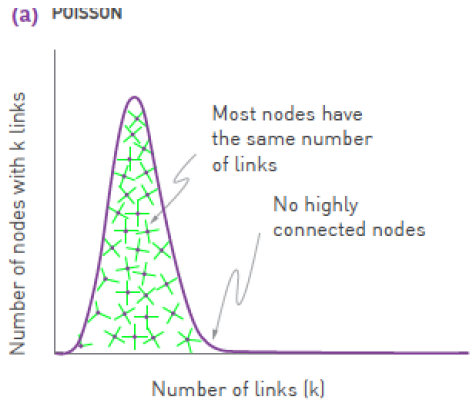
| POPULARITY | QUALITY | MIXED MODEL |
| --- | --- | --- |
| We want to be associated with popular people, ideas, items, thus further increasing their popularity, irrespective of any objective measurable characteristics | We evaluate people and everything else based on objective quality criteria, so higher quality nodes will naturally attract more attention faster | Among nodes of similar attributes, those that reach critical mass first will become 'star' with many friends and followers ('halo effect') |
| also known as 'the rich get richer' | also known as 'the good get better' | may be impossible to predict who will become a star, even if quality matters |

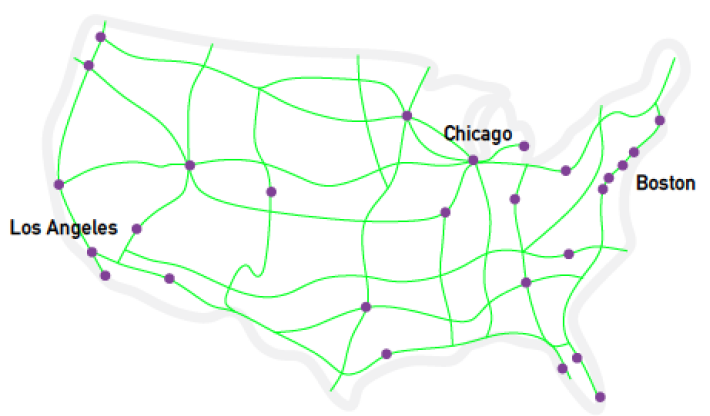
# Core-Periphery Structures

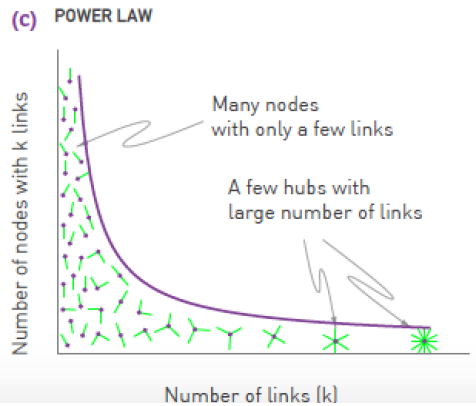
- A useful and relatively simple metric of the degree to which a social network is centralized or decentralized, is the centralization measure (usually normalized such that it takes values between 0 and 1)
  - It is based on calculating the differences in degrees between nodes; a network that greatly depends on 1-2 highly connected nodes (as a result for example of preferential attachment) will exhibit greater differences in degree centrality between nodes
  - Centralized structures can perform better at some tasks (like team-based problem-solving requiring coordination), but are more prone to failure if key players disconnect
- In addition to centralization, many large groups and online communities have a core of densely connected users that are critical for connecting a much larger periphery
  - Cores can be identified visually, or by examining the location of high-degree nodes and their joint degree distributions (do high-degree nodes tend to connect to other high-degree nodes?)
  - Bow-tie analysis, famously used to analyze the structure of the Web, can also be used to distinguish between the core and other, more peripheral elements in a network

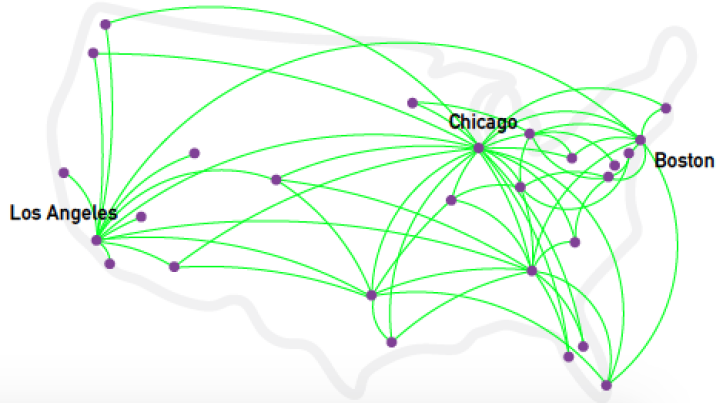


# Social Network Characteristic

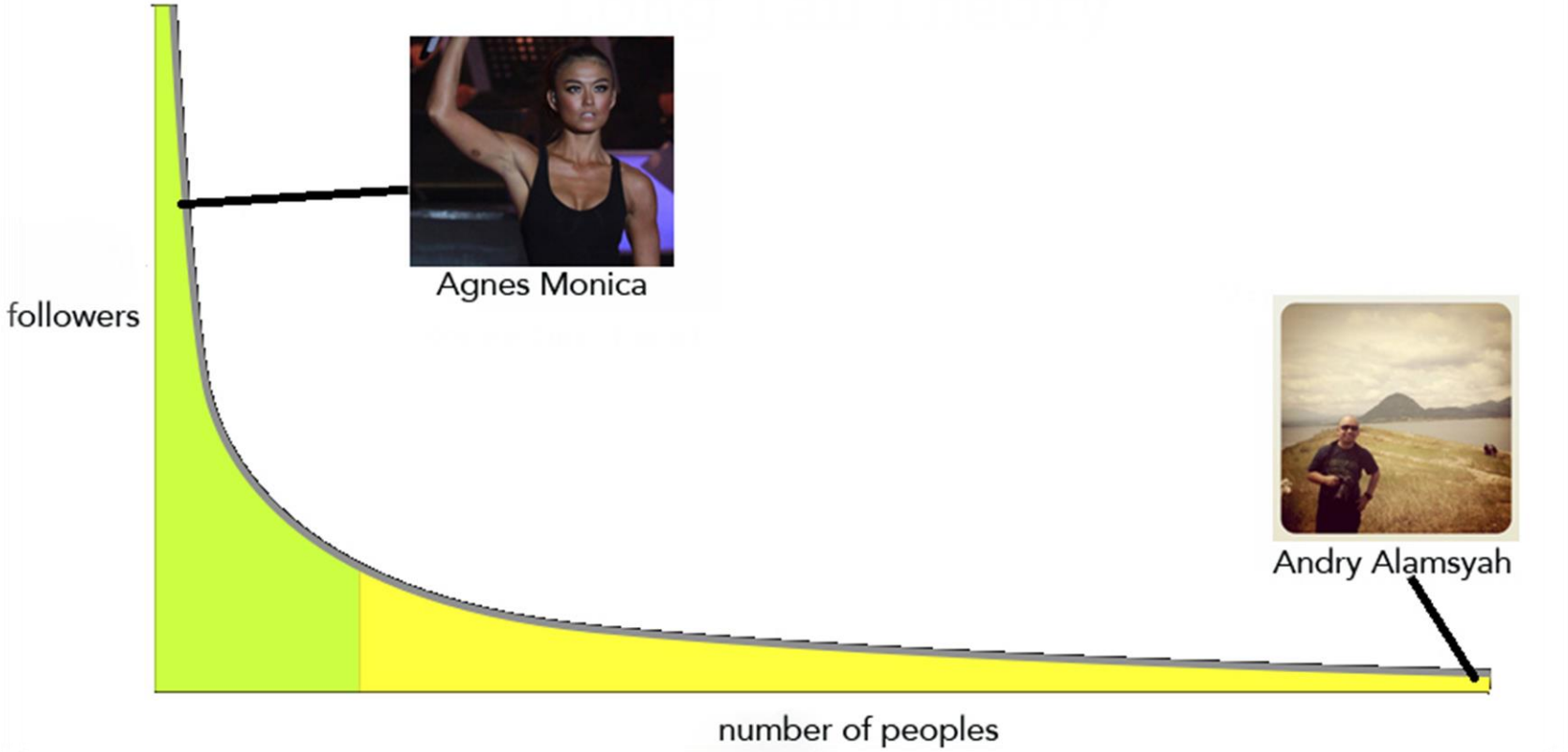


# Social Network Characteristic: Power Law




power law distributions

# Practice: Constructing Interaction Graph

# 3 Collecting Social Media Data

Social Media Analytic Workshop

Telkom University

DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Learning Outline

- **Introduction**
- Collecting Data from Twitter
- Constructing Online Interaction Graph
- Collecting Data from Social Forum

# Collecting Data from Social Media

- First, the most direct way is to download the data from databases on the web servers.
- The second method is collecting data through application programming interfaces (APIs).
- The third method is web scraping. This method is particularly useful for those web-sites that do not provide APIs.

# Ethical issues

- Harvesting social media data may lead to an unwanted and unforeseen exposure of private data for participants .Large quantities of personal information on social media.
- Informed consent
- Participant anonymity

# Learning Outline

- Handshake with Social Media Data
- **Collecting Data from Twitter**
- Constructing Online Interaction Graph
- Collecting Data from Social Forum

# Practice: Collecting Twitter Data Using R

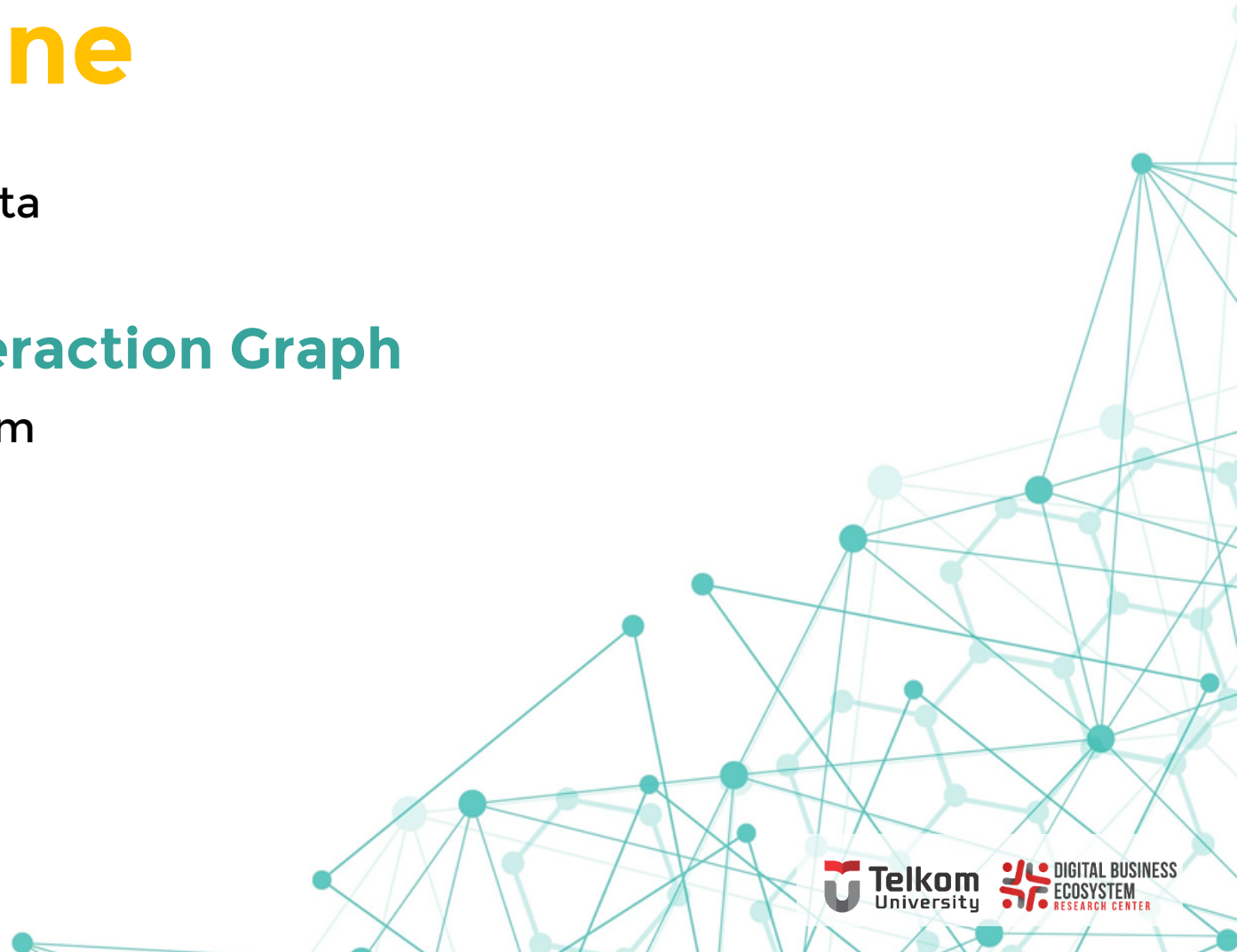

Practice Link: https://github.com/sma-workshop/r

# Practice: Collecting Twitter Data Using Python


Practice Link: https://github.com/sma-workshop/python

# Learning Outline

- Handshake with Social Media Data
- Collecting Data from Twitter
- **Constructing Online Interaction Graph**
- Collecting Data from Social Forum

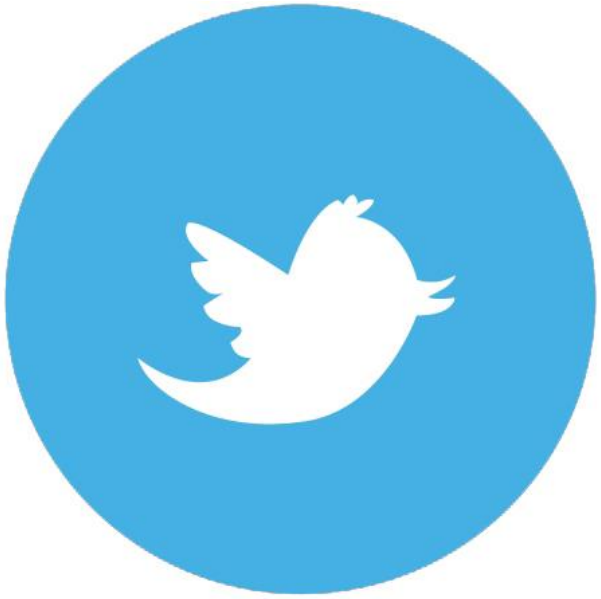


Telkom University
DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Build the Dataset

| screen_name<br>=<br>Source | reply_to_screen<br>=<br>Target |
|---|---|
| Dian | Rizqy |
| Dito | Rizqy |
| Dito | Dian |
| | |
| ... | ... |

# Practice: Constructing Online Interaction Graph



# Learning Outline

- Handshake with Social Media Data
- Collecting Data from Twitter
- Constructing Online Interaction Graph
- **Collecting Data from Social Forum**

# Practice: Collecting Review Data

# 4 Text Mining

Social Media Analytic Workshop

Telkom University

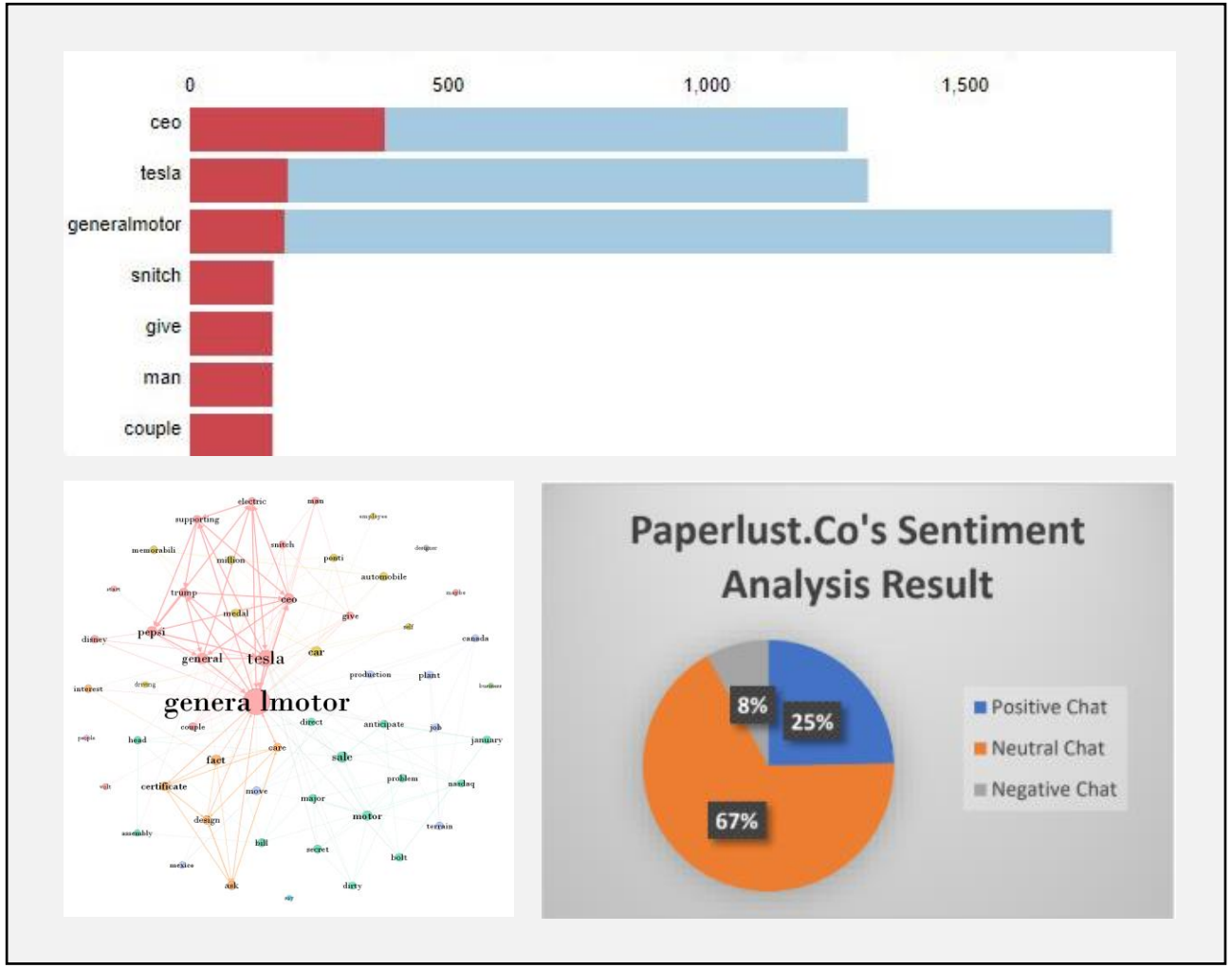
DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Learning Outline

- **Introduction to Text Mining**
- Text Preprocessing
- Text Classification: Sentiment Analysis
- Topic Modelling
- Text Network Analysis

# Text Mining

- According to Wikipedia, “Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text.”
- Text mining combines notions of statistics, linguistics, and machine learning to create models that learn from training data and can predict results on new information based on their previous experience.

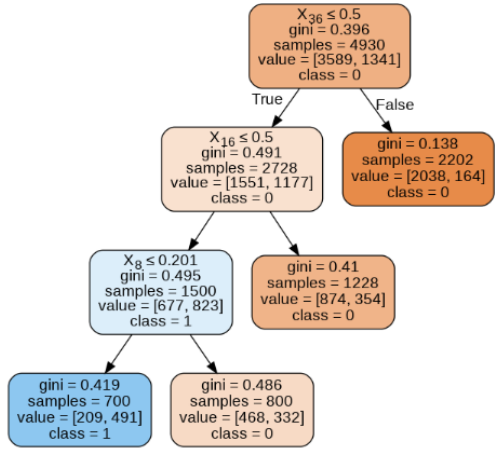


# Data Mining and Text Mining

## Data Mining

| customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService |
|---|---|---|---|---|---|---|---|---|
| 7590-VHVEG | Female | No | Yes | No | 1 | No | No phone service | DSL |
| 5575-GNVDE | Male | No | No | No | 34 | Yes | No | DSL |
| 3668-QPYBK | Male | No | No | No | 2 | Yes | No | DSL |
| 7795-CFOCW | Male | No | No | No | 45 | No | No phone service | DSL |
| 9237-HQITU | Female | No | No | No | 2 | Yes | No | Fiber optic |
| 9305-CDSKC | Female | No | No | No | 8 | Yes | Yes | Fiber optic |
| 1452-KIOVK | Male | No | No | Yes | 22 | Yes | Yes | Fiber optic |
| 6713-OKOMC | Female | No | No | No | 10 | No | No phone service | DSL |
| 7892-POOKP | Female | No | Yes | No | 28 | Yes | Yes | Fiber optic |
| 6388-TABGU | Male | No | No | Yes | 62 | Yes | No | DSL |
| 9763-GRSKD | Male | No | Yes | Yes | 13 | Yes | No | DSL |
| 7469-LKBCI | Male | No | No | No | 16 | Yes | No | No |
| 8091-TTVAX | Male | No | Yes | No | 58 | Yes | Yes | Fiber optic |
| 0280-XJGEX | Male | No | No | No | 49 | Yes | Yes | Fiber optic |
| 5129-JLPIS | Male | No | No | No | 25 | Yes | No | Fiber optic |

Output




## Text Mining

Sebelum ricuh, perwakilan mahasiswa sudah diterima pimpinan DPRD untuk menyampaikan aspirasinya. Aparat keamanan kemudian mengimbau massa agar tak mendesak masuk gedung dewan. Namun imbauan itu tak diindahkan mahasiswa.

Bersamaan dengan lemparan botol, batu, sepatu dan barang-barang lainnya, mahasiswa yang berada depan pagar merangsek masuk. Pagar yang tadinya menyekat mahasiswa dengan aparat keamanan akhirnya jebol.

Output

Important Keyword:
Ricuh
Mahasiswa
DPR
Aparat
Desak
Masuk

# Learning Outline

- Introduction to Text Mining
- **Text Preprocessing**
- Text Classification: Sentiment Analysis
- Topic Modelling
- Text Network Analysis

# Everything was fine, until you meet Indonesian Language

**Kandiya R. Maulidita** @R_IndyJKT48 · 8h

Woi anak orang yang baik nya tulus amat! Habedeee nandskiiiiii mi amor😘😘 Sampe setaun lbh ini gue kenal **lu**, gue gapernah ngeliat **lu** marah2 ato sebel sm org lama bgtt... Sumpah gw gangerti **lu bisa** setulus itu baiknya😍 I LOP UU NANDSKI!! SUKSES SELALU YAW❤️❤️ #L16htenUpNandays

**siam** 🔞 @mimiamiamia95 · 3m

O gitu caranya banya followers bikin GA ipon, pas udah banyak yang pollow GAnya gajadi diumumin deh atau bilangnya udah ada yang menang hhhhhh Bisa saja **klean** ini ngibulnya 😍😍

Telkom University | DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

**siam** 18 @mimiamiamia95 · 3m

O gitu caranya banya followers bikin GA ipon, pas udah banyak yang pollow GAnya gajadi diumumin deh atau bilangnya udah ada yang menang hhhhhh Bisa saja **klean** ini ngibulnya 😊😊

| Pre-Precessing Type | Result |
|---|---|
| Tokenization | O gitu caranya banya followers bikin GA ipon, pas sudah banyak yang pollow GAnya gajadi diumumin deh atau bilangnya sudah ada yang menang hhhhhh Bisa saja klean ini ngibulnya |
| Slang word | O gitu caranya banyak followers bikin GA iphone, pas sudah banyak yang follow GAnya tidak jadi diumumin deh atau bilangnya sudah ada yang menang hhhhhh Bisa saja kalian ini ngibulnya |
| Stemming | O gitu cara banyak followers bikin GA iphone, pas sudah banyak yang follow GA tidak jadi umum deh atau bilang sudah ada yang menang hhhhhh Bisa saja kalian ini ngibul |
| Lemmatization | O gitu cara banyak followers buat GA iphone, pas sudah banyak yang follow GA tidak jadi umum deh atau sebut sudah ada yang menang hhhhhh Bisa saja kalian ini tipu |
| Stop word | cara banyak followers GA iphone, banyak follow GA tidak umum sebut menang kalian tipu. |

# Practice: Pre-Processing Text Data Using R


Practice Link: https://github.com/sma-workshop/r

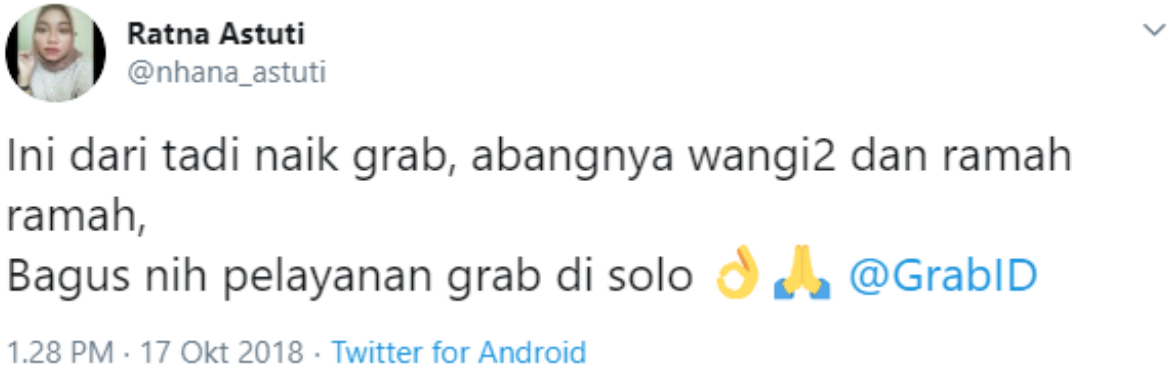
# Learning Outline

- Introduction to Text Mining
- Text Preprocessing
- **Text Classification: Sentiment Analysis**
- Topic Modelling
- Text Network Analysis

Telkom University
DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

# Text Classification

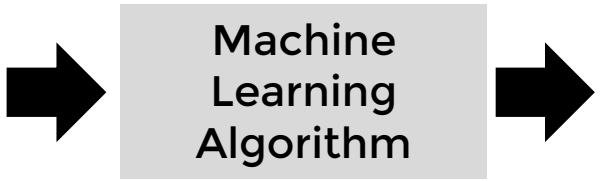
Text classification also known as text tagging or text categorization is the process of categorizing text into organized groups. Some of the most common examples and use cases for automatic text classification is **Sentiment Analysis.**




# Create Sentiment Analysis Model

It is using Machine Learning Principle

| Tweets | Sentiment |
|---|---|
| kartu yg sdh sukses diregistrasi ttp di reject saat bayar via grabpay?mslh baru timbul semingguan terakhir | Negative |
| halo Grab kenapa ya lokasi ko ga kedetect dari tadi? sy mo order ga bisa | Negative |
| pengemudi payah..ambil penumpang dr bandara sll cancel! | Negative |
| apa pembatalan sama dengan order yang tidak di ambil? | Negative |
| klo mau ganti no rek cimb kemana ya? troubleshootnya lambat menanggapinya | Negative |
| SECEPAT datangnya GRAB\ud83d\02 Buruan pake !! Radiovenusmks | Neutral |
| min mau top up grabpay di alfamart sistemnya eror tuh | Negative |




| Tweets | Sentiment | Predicted Sentiment |
|---|---|---|
| tidak bisa dipakai sama sekali saldonya. Mohon di kroscek ya | | Negative |
| aplikasi grab kok gbsa ya? Keluar sendiri pas d klik | | Negative |
| halo kenapa applikasi grab saya selalu ketutup sendiri stlh order grabfood ya? Saya jd tdk bisa hub drivernya | | Negative |
| I can't access Grabchat. May I know why? | | Negative |
| kalo pesen grabfood..kdg di app open..tp aslinya close. driver nya suka kayak bete gt ditelfon seolah olah kt yg salah ya haha | | Negative |
| halo dari kemarin kalo jam segini mau pesen grab selalu \failed to contact the Grab server\ padahal kalo pagi ga ada masalah | | Negative |

# Practice: Sentiment Analysis Using R


Practice Link: https://github.com/sma-workshop/r

# Practice: Sentiment Analysis Using Python

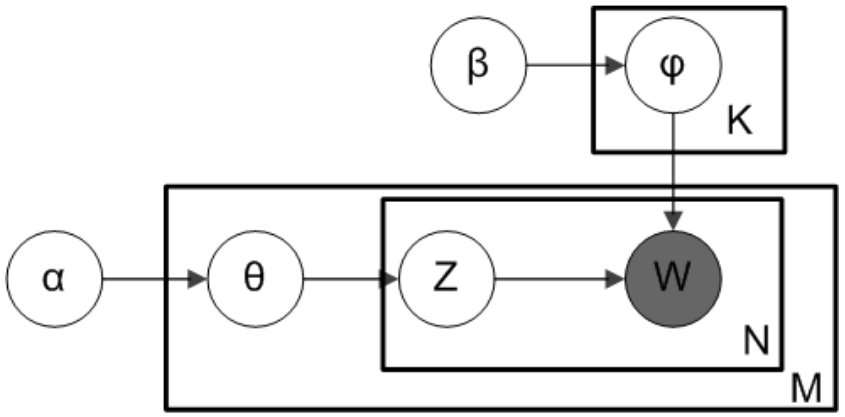

Practice Link: https://github.com/sma-workshop/python

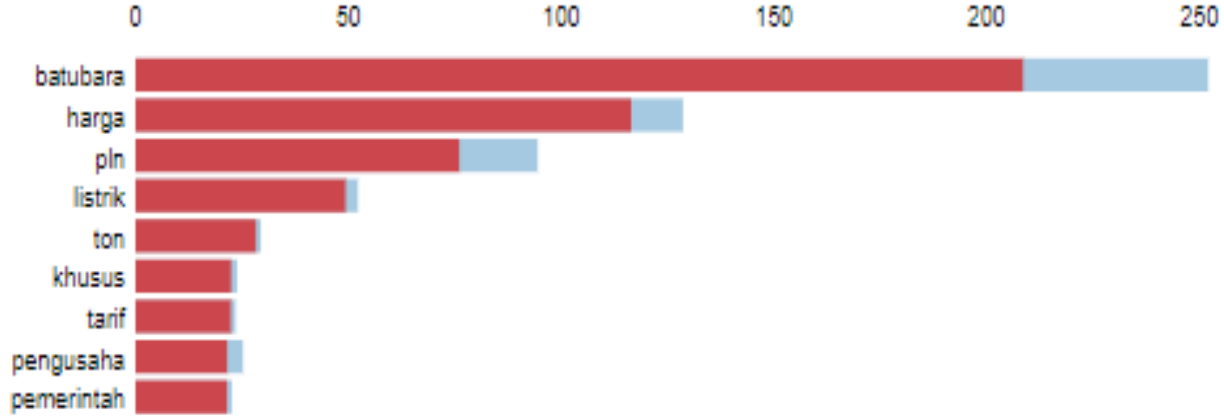
# Learning Outline

- Introduction to Text Mining
- Text Preprocessing
- Text Classification: Sentiment Analysis
- **Topic Modelling**
- Text Network Analysis

# Topic Modelling

- Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents..
- LDA is easily the most popular (and typically most effective) topic modeling technique out there.



# Practice: LDA Topic Modelling Using Python


Practice Link: https://github.com/sma-workshop/python

# Practice: LDA Topic Modelling Using R


Practice Link: https://github.com/sma-workshop/r

# Learning Outline

- Introduction to Text Mining
- Text Preprocessing
- Text Classification: Sentiment Analysis
- Topic Modelling
- **Text Network Analysis**

# Text Network Analysis

- The analysis of text describing various kinds of "computer support solution" that allows the analyst to "extract networks of concepts" from text
- The assumption underlying the emergence text network is a relationship that is contained in every word so that it produces a pattern that can be analyzed

# Practice: Text Network Analysis

WORDij

# Thank You