Roll No: | | | | | | | | | | | |

PRESIDENCY UNIVERSITY

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956
Approved by AICTE, New Delhi | Approved By BCI
Bengaluru

Mid Term Examinations - October 2026

Date: 05/10/2026 **Time:** 11:45 AM - 01:15 PM

Course Code: CSD3445 **Course Name:** Social Media Analytics

Semester: Seventh Semester **Max. Marks:** 50 **Weightage:** 50%

CO - Levels	CO1	CO2
Marks	34	56

PART A: Answer Following Questions. 10 M

Q.N.	Questions	M	CO	BT
1	**Definition.** The social media ecosystem is the **interconnected environment of users, platforms, content, technologies, businesses and governing rules** in which social information is created, shared, consumed and monetised. The word "ecosystem" stresses that the parts depend on one another: users produce content on platforms, algorithms distribute it, advertisers fund the platforms, and policies regulate everybody. **Key component** / **Role in the ecosystem** / **Examples** Users / actors — Create, consume and share content; form relationships — Individuals, influencers, brands, media houses Platforms — Host profiles, content and interactions — Facebook, Instagram, X, LinkedIn, YouTube User generated content — Main asset and main source of data — Posts, videos, reviews, comments Relationships and interactions — Form the social graph; carry information — Follow, friend, like, share, mention Algorithms and recommendation systems — Decide what each user sees — Feed ranking, suggestions Advertisers, APIs and policies — Fund the platforms, expose data, regulate behaviour — Ad networks, platform APIs, moderation rules	2	CO1	BT1
2	*Differentiate between reach and impressions as social media metrics.* **Basis** / **Reach** / **Impressions** Meaning — Number of **unique** users who saw the content at least once — **Total number of times** the content was displayed, including repeat views by the same user How it is counted — Each user is counted only once — Every display is counted; one user can be counted many times Relationship — Reach ≤ impressions — Impressions = reach × average frequency What it indicates — Size of the audience actually touched (breadth) — Volume of exposure (intensity or repetition) Example — A post seen by 1,000 different users: reach = 1,000 — The same post displayed 3,500 times in total: impressions = 3,500 *Frequency = Impressions ÷ Reach = 3,500 ÷ 1,000 = 3.5 views per user* Reach is used to judge how many different people a campaign has reached (awareness), whereas impressions are used to judge how heavily the content was delivered (exposure), for instance in advertising billing.	2	CO1	BT2
3	*Define degree centrality of a node. How is it normalised for a network of n nodes?* **Degree centrality.** The degree centrality of a node is the **number of edges incident on it**, that is, its number of direct connections: $C_d(v) = d(v)$ It measures how popular or active a node is; a node with many direct contacts can reach and influence many others immediately. In a directed network, *in-degree* (number of incoming edges, e.g., followers) measures popularity or prestige, and *out-degree* (outgoing edges, e.g., accounts followed) measures sociability. **Normalisation.** In a simple graph with n nodes a node can be connected to at most (n − 1) other nodes, so *Normalised* $C_d(v) = d(v) \div (n - 1)$ The value lies between 0 and 1 and allows comparison between networks of different sizes. **Example:** a node with 3 neighbours in a network of 6 nodes has normalised degree centrality 3 ÷ 5 = 0.6.	2	CO2	BT1

ANSWER KEY

<table>
<tr>
<td>4</td>
<td>Community. A community is a group of nodes that are densely connected to one another and only sparsely connected to the rest of the network – for example classmates, colleagues or the members of a hobby group. Formally, the density of edges inside the group is clearly higher than the density of edges between the group and the outside.
<table>
<tr><th>Approach</th><th>Basic idea</th></tr>
<tr><td>Girvan–Newman (divisive)</td><td>Repeatedly remove the edge of highest edge betweenness; the components that appear are communities</td></tr>
<tr><td>Louvain / modularity maximisation</td><td>Greedily move nodes between communities and aggregate communities to maximise modularity</td></tr>
<tr><td>Label propagation</td><td>Each node adopts the label most common among its neighbours until labels stabilise</td></tr>
<tr><td>Clique percolation</td><td>Communities are unions of adjacent k-cliques; allows overlapping communities</td></tr>
<tr><td>Hierarchical / spectral clustering</td><td>Cluster nodes by similarity of connections or by eigenvectors of the graph Laplacian</td></tr>
</table>
</td>
<td>2</td>
<td>CO2</td>
<td>BT1</td>
</tr>
<tr>
<td>5</td>
<td>Link analysis. Link analysis is the study of the links (edges) between entities in a network in order to discover the importance of nodes, the structure of relationships and hidden or missing connections. Instead of looking only at the attributes of individual nodes, it uses the pattern of connections. Its main techniques are link-based ranking (PageRank, HITS), link prediction and relationship (link) mining.
<table>
<tr><th>Application</th><th>How link analysis is used</th></tr>
<tr><td>Friend / follower recommendation</td><td>Link prediction scores unconnected pairs (e.g., by common neighbours) and suggests the best candidates</td></tr>
<tr><td>Identifying influential users</td><td>Link-based ranking (PageRank, in-degree) finds users whose links carry most authority</td></tr>
<tr><td>Spam, bot and fraud detection</td><td>Unusual link patterns, such as tightly knit groups following one another, reveal fake accounts</td></tr>
<tr><td>Community detection</td><td>Densely linked groups are extracted from the link structure</td></tr>
</table>
</td>
<td>2</td>
<td>CO2</td>
<td>BT2</td>
</tr>
</table>

PART B : Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**

ANSWER KEY

(a) User behaviour analytics and its role [5 M]

Meaning. User behaviour analytics (UBA) is the process of **collecting, measuring and interpreting data about what users do** on a platform – how they log in, what they post, like, comment on, share, follow, watch and click, and when and for how long – in order to understand their patterns, preferences and intentions and to predict what they are likely to do next. Because people leave a digital footprint of nearly everything they do on social media, behaviour can be studied at very large scale and at very fine detail.

Data used in UBA.

Category of data	Examples
Activity data	Logins, posts, comments, likes, shares, follows, messages
Consumption data	Views, dwell time, watch time, clicks, scroll depth, searches
Social data	Friends, followers, group memberships, interaction partners
Context data	Device, time of day, location, language
Content data	Text, images and videos created by the user (for sentiment and topic analysis)

The UBA process. Behaviour data are analysed in a pipeline of six stages:

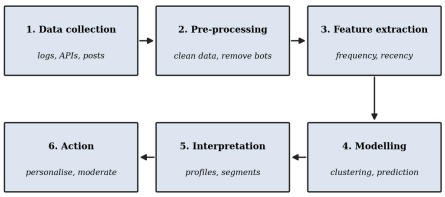


Fig. 1: Process of user behaviour analytics

- **Data collection** from platform logs and APIs, respecting user consent and privacy.
- **Pre-processing:** cleaning, removing spam and automated (bot) activity, resolving duplicates.
- **Feature extraction:** activity frequency, recency, content type, session length, time-of-day pattern, network position.
- **Modelling and analysis:** clustering and segmentation, classification and prediction (churn, interest, purchase), sequence and time-series analysis, anomaly detection.
- **Insight and interpretation:** user profiles, segments, behavioural trends.
- **Action:** personalised recommendations, campaigns, moderation and product decisions.

Role of UBA in understanding users.

Role	How UBA helps	Example
Understanding interests and intent	Reveals what users like and what they are likely to do next	Topics a user follows and watches
Personalisation	Tailors feed, recommendations and notifications	"Suggested for you" videos
Segmentation	Groups users with similar behaviour	Lurkers, contributors, power users
Prediction	Predicts churn, engagement or purchase	Users whose activity is declining
Content and campaign optimisation	Shows which content types and timings work	Best posting time for an audience
Detecting abnormal behaviour	Flags bots, fake accounts, account take-over and abuse	Hundreds of identical posts per minute
Community health	Monitors participation and satisfaction	Falling reply rate in a forum

6 | 10 | CO2 | BT2

Ethical note. Because UBA works on personal data, it must be carried out with user consent, data minimisation and anonymisation, and with attention to privacy laws and to bias in the models.

(b) Individual versus collective behaviour analysis [5 M]

Social media mining distinguishes two levels of behaviour analytics. **Individual behaviour analysis** studies the behaviour of a single user – what the user does, why, and what the user will do next. **Collective behaviour analysis** studies the behaviour that **emerges when many individuals act and interact together**, which is often different from, and not simply the sum of, what each individual would do alone.

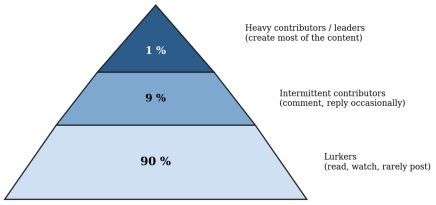
Basis	Individual behaviour analysis	Collective behaviour analysis
Unit of analysis	A single user	A crowd, a community or the whole population of users
Central question	What does this user do, and what will he or she do next?	What pattern emerges when many users act together?
Data used	The user's profile, activity history, content and social ties	Aggregated activity of many users: volume of hashtags, shares, cascades, group activity
Typical techniques	Classification, regression, user profiling, recommendation, anomaly detection	Trend detection, information-diffusion and cascade models, network and population-level models, time-series analysis
Examples	Predicting whether a user will buy a product; recommending videos; detecting a compromised account	A viral hashtag challenge; trending topics during a match; herd behaviour in a rumour or panic; collective mood about a product launch
Outcome	Personalisation and individual-level decisions	Understanding social phenomena; population-level and policy decisions

Illustration. Suppose thousands of users share the same challenge video within a few days. *Individual analysis* asks why a particular user joined – because friends did, because of interest in the cause, or because of a personal habit of joining challenges. *Collective analysis* asks how fast the challenge spread, how many people took part at its peak and which communities carried it.

Link between the two. Collective behaviour arises from individual decisions and their interactions, so good models combine

ANSWER KEY

PTO (Please Turn Over)

7	**(a) Characteristics of online community behaviour [5 M]** An **online community** is a group of people who interact repeatedly through an online platform – a forum, a group, a subreddit or a comment section – because of a shared interest, purpose or identity. Its behaviour has the following characteristics: **Characteristic** | **Explanation** 1. Shared purpose or interest | Members gather around a common topic, goal or identity (a programming language, a hobby, a cause). This purpose drives what is posted and discussed. 2. Norms and rules | Explicit guidelines and unwritten etiquette define acceptable behaviour; moderators enforce them and newcomers learn them by observation. 3. Roles and status | Members take roles such as leaders/moderators, experts, regular contributors, occasional participants and newcomers; status is earned by contribution and reputation. 4. Reciprocity, trust and cohesion | Members help each other and return favours; repeated interaction builds trust, identity and a sense of belonging. 5. Member life cycle | Members move from newcomer to regular to leader, and some leave (churn). The health of a community depends on attracting newcomers and retaining active members. 6. Unequal participation | A small fraction of members generates most of the content – see below. **Participation inequality (the 90–9–1 rule).** In most online communities participation is highly unequal and follows a skewed, power-law-like pattern. As a rule of thumb (popularised by usability researcher Jakob Nielsen), about **90 % of users are lurkers**, who read, watch and search but never contribute; about **9 % contribute occasionally**; and about **1 % are heavy contributors who create the large majority of the content**. The exact proportions differ between communities, but the shape – a wide base of silent readers and a small core of active creators – is very common. In Wikipedia, for example, a small core of editors makes a very large share of all edits.  *Fig. 1: Participation inequality (90–9–1 rule)* **Group** | **Approx. share** | **Behaviour** | **Importance to the community** Lurkers | ~90 % | Read, watch, search; rarely or never post | Form the audience; a large pool of potential contributors Intermittent contributors | ~9 % | Reply, comment or post from time to time | Keep the conversation going; provide diversity of views Heavy contributors / leaders | ~1 % | Create most content, answer questions, moderate | Set the tone, provide expertise and hold the community together The *reader-to-leader framework* (Preece and Shneiderman) describes how members can progress from readers to contributors, collaborators and leaders as they gain experience and trust. **Implication for analysis:** the visible content comes from a small minority and therefore does not represent the opinions of the silent majority; analysts must take this bias into account. **(b) Contribution of participation patterns to social intelligence [5 M]** Social intelligence is the ability to turn social data into insight for decisions. Participation patterns are among its richest sources because they show *who* does *what*, *how often* and *with what effect* in a community. **Participation pattern observed** | **Social intelligence obtained** | **How it is used** Who posts, answers and moderates; who is replied to most | Identification of leaders, experts and influencers | Engage experts and ambassadors; target information campaigns Activity level, retention, reply time, ratio of lurkers to contributors | Measure of community health and growth | Detect declining communities early; set targets Topics that attract high participation | Interests, needs and concerns of members | Guide product, content and policy decisions Declining activity of a member or group | Early warning of disengagement (churn) | Timely retention actions Conflicts, spam and abusive behaviour | Signals for moderation and safety | Improve moderation and community rules Conversion of lurkers to contributors | Understanding of what motivates participation | Design incentives, onboarding and recognition Taken together, participation analysis moves an organisation from *data* (counts of posts) to *insight* (who matters, what members want, whether the community is healthy) and finally to *decisions*, which is the foundation of social intelligence.	10	CO2	BT2

PART C: Answer Any 1 Following Questions: 10 M

Q.N.	Questions	M	CO	BT

ANSWER KEY

(a) Classification of social media metrics [4 M]

A **social media metric** is a quantitative measure of performance on a platform. Metrics are chosen according to the business objective and are classified by the stage of the audience journey they measure – the *marketing funnel* of awareness, engagement and conversion. Each stage has fewer people than the one before it.

Category	Objective	Typical metrics (examples)	Question answered
Awareness metrics	Measure how many people are exposed to the brand or content	Impressions, reach, follower growth, video views, brand mentions, share of voice	How many people know about us?
Engagement metrics	Measure how actively the audience interacts	Likes, comments, shares, saves, engagement rate, average watch time	Do people care enough to interact?
Conversion metrics	Measure business outcomes that follow interaction	Clicks, click-through rate (CTR), conversion rate, leads, sign-ups, cost per click (CPC), cost per conversion, return on investment	Do interactions turn into results?

Analysts compare the numbers at each stage of the funnel to find where the audience is lost (for example many impressions but a low CTR points to unattractive content).

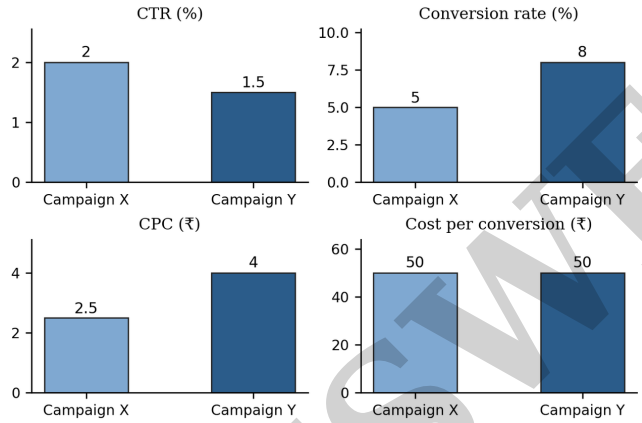
(b)(i) Computation of the measures [4 M]

Campaign X: spend = ₹20,000; impressions = 4,00,000; clicks = 8,000; conversions = 400.

Campaign Y: spend = ₹30,000; impressions = 5,00,000; clicks = 7,500; conversions = 600.

Measure	Formula	Campaign X	Campaign Y
CTR	(clicks ÷ impressions) × 100	8,000 ÷ 4,00,000 × 100 = **2.0 %**	7,500 ÷ 5,00,000 × 100 = **1.5 %**
Conversion rate	(conversions ÷ clicks) × 100	400 ÷ 8,000 × 100 = **5.0 %**	600 ÷ 7,500 × 100 = **8.0 %**
CPC	spend ÷ clicks	₹20,000 ÷ 8,000 = **₹2.50**	₹30,000 ÷ 7,500 = **₹4.00**
Cost per conversion	spend ÷ conversions	₹20,000 ÷ 400 = **₹50**	₹30,000 ÷ 600 = **₹50**

8 — 10 — CO1 — BT3



(b)(ii) Comparison and recommendation [2 M]

- **Traffic efficiency – X is better.** X earns more clicks for every 100 impressions (CTR 2.0 % against 1.5 %), so its creative attracts attention better, and each click costs less (₹2.50 against ₹4.00).
- **Visitor quality – Y is better.** Y converts 8 % of its visitors compared with 5 % for X, so its audience or landing offer is more relevant, and it produced more conversions in total (600 against 400).
- **Cost efficiency – equal.** Both campaigns cost ₹50 per conversion, so neither wastes more money per result.

Recommendation. If the objective is **sales or sign-ups (conversions)**, choose **Campaign Y**: at the same cost per conversion it delivers 50 % more conversions and higher-quality traffic, so more budget can be placed on it with confidence. If the objective is **cheap traffic or awareness**, choose Campaign X because of its higher CTR and lower CPC. A good analyst would also combine the strengths of both – X's creative with Y's audience targeting – and test the change through an A/B experiment.

ANSWER KEY

(a) Techniques of social data visualization [5 M]

Social data visualization is the graphical presentation of social media data so that patterns, trends, relationships and outliers can be seen quickly by analysts and decision makers. Because social data are large, noisy and network-shaped, the choice of chart must match the *type of data* and the *question being asked*.

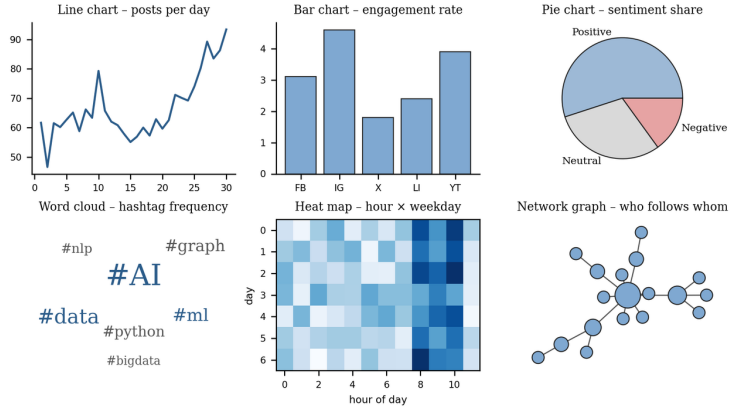


Fig. 1: Common visualizations for social data (illustrative data)

Technique	Description	Suitable when	Social media example
1. Bar / column chart	Bars whose length is proportional to a value	Comparing a measure across a few discrete categories; ranking	Engagement rate of five platforms
2. Line chart (time series)	Points joined in time order	Showing trends, seasonality and spikes over time	Daily post volume; follower growth
3. Pie / donut chart	Circle divided into slices	Showing parts of a whole with very few categories	Share of positive, neutral and negative sentiment
4. Word cloud	Words sized according to frequency	Quick, informal overview of frequent terms in text	Frequent hashtags or words in tweets
5. Heat map	Matrix whose cells are coloured by value	Two categorical dimensions with intensity; finding peak periods	Activity by hour and weekday
6. Network (node–link) graph	Nodes and edges with size and colour encodings	Showing relationships, communities and influential nodes	Follower network, retweet network
7. Scatter plot / map (optional)	Points on two axes or on a map	Relationship between two variables; location-based data	Followers versus engagement; geo-tagged posts

Good practice. Match the chart to the data type; label axes and units; avoid three-dimensional effects and pie charts with many slices; use colour purposefully and accessibly; limit clutter; and keep the story simple. Common tools are Matplotlib/Seaborn and Plotly in Python, Tableau and Power BI for dashboards, and Gephi or NetworkX for network graphs.

(b) Choice of visualization for the given tasks [5 M]

Task	Most suitable visualization	Justification
(i) Daily post volume over one month	**Line chart** (time series), optionally with a moving average	Data are ordered in time; the line makes trend, peaks and dips across 30 consecutive days immediately visible
(ii) Most frequent hashtags in a discussion	**Word cloud** for a quick overview, or a **horizontal bar chart** of the top-N hashtags	The data are frequencies of text items; a word cloud gives an instant impression, while bars give the exact ranking and counts
(iii) Who-follows-whom among 50 users	**Network (node–link) graph** with a force-directed layout; node size by degree	The data are relationships; only a graph shows connections, clusters and hubs. With 50 nodes the graph is still readable
(iv) Engagement rate across five platforms	**Bar chart**	Compares one numeric measure across a few separate categories; bar length can be compared accurately
(v) Activity by hour of day and day of week	**Heat map** (24 × 7 matrix; colour = activity)	Two categorical dimensions plus an intensity; colour reveals peak days and hours (for example weekday evenings)

9		10	CO1	BT3

ANSWER KEY

PART D: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**

10

(a) Centrality measures [3 M]

Centrality measures answer the question "which nodes are most important?". Because importance can mean different things – having many contacts, being close to everybody, or standing between others – there are several measures. Let d(v) be the degree of v, l(v,u) the length of the shortest path between v and u, σst the number of shortest paths from s to t, and σst(v) the number of those paths that pass through v.

Centrality	What it measures	Formula	Social meaning
Degree	Number of direct connections of a node	$d(v) \div (n-1)$	Popularity or activity; local influence
Closeness	Inverse of the average shortest-path distance from the node to all others	$(n-1) \div \Sigma_{u \neq v}\, l(v,u)$	How quickly the node can reach or hear from everyone; efficient spreader of information
Betweenness	Fraction of shortest paths between other pairs that pass through the node	$\Sigma_{s \neq v \neq t}\, \sigma st(v) \div \sigma st$	Broker or gatekeeper; control over the flow between groups

(b) Degree and closeness centrality [5 M]

n = 6, so n − 1 = 5. Adjacency list: 1: {2}; 2: {1, 3, 4}; 3: {2, 4}; 4: {2, 3, 5}; 5: {4, 6}; 6: {5}.

Step 1 – Degree centrality. d(1) = 1, d(2) = 3, d(3) = 2, d(4) = 3, d(5) = 2, d(6) = 1 (sum = 12 = 2m). Normalised value = d ÷ 5.

Step 2 – Shortest-path distance matrix (by breadth-first search from each node):

From \ To	1	2	3	4	5	6	Σ distances
1	0	1	2	2	3	4	12
2	1	0	1	1	2	3	8
3	2	1	0	1	2	3	9
4	2	1	1	0	1	2	7
5	3	2	2	1	0	1	9
6	4	3	3	2	1	0	13

Step 3 – Closeness = (n − 1) ÷ Σ distances = 5 ÷ Σ distances. **Step 4 – Results:**

Node	Degree d	Normalised degree (d ÷ 5)	Σ distances	Normalised closeness (5 ÷ Σ)
1	1	0.20	12	0.417
2	3	0.60	8	0.625
3	2	0.40	9	0.556
4	3	0.60	7	**0.714**
5	2	0.40	9	0.556
6	1	0.20	13	0.385

Nodes 2 and 4 share the highest degree centrality (0.60), but node 4 has the highest closeness (0.714), because it lies in the middle of the network; the end nodes 1 and 6 are the most peripheral.

(c) Betweenness centrality of node 4 [2 M]

Node 4 lies on the shortest path between every node of {1, 2, 3} and every node of {5, 6}. Each such pair has exactly one shortest path (σst = 1):

Pair (s, t)	Shortest path	σst	σst(4)	Contribution
1 – 5	1 – 2 – 4 – 5	1	1	1
1 – 6	1 – 2 – 4 – 5 – 6	1	1	1
2 – 5	2 – 4 – 5	1	1	1
2 – 6	2 – 4 – 5 – 6	1	1	1
3 – 5	3 – 4 – 5	1	1	1
3 – 6	3 – 4 – 5 – 6	1	1	1

$C_b(4) = 6$ (unnormalised);

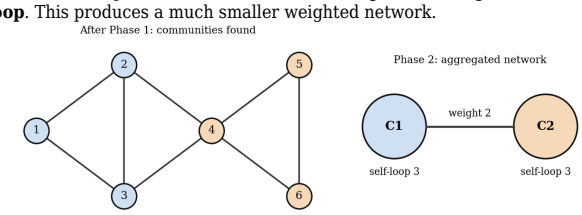
normalised = 6 ÷ [(n − 1)(n − 2) ÷ 2] = 6 ÷ 10 = 0.6

Pairs inside {1, 2, 3} (for example 1–3, which uses 1–2–3) do not pass through node 4. For comparison, $C_b(2) = 4$ (pairs of node 1 with 3, 4, 5, 6), $C_b(5) = 4$ (pairs of node 6 with 1, 2, 3, 4), and $C_b(1) = C_b(3) = C_b(6) = 0$.

Why node 4 is the most critical. Node 4 has the **highest betweenness (6)**, the **highest closeness (0.714)** and shares the highest degree (0.60). It is the only connection between the part {1, 2, 3} and the tail {5, 6}: if node 4 is removed, nodes 5 and 6 are cut off from the rest of the network and all information flow between the two parts stops. Node 4 is therefore a cut vertex and the main broker of the network.

10 | CO1 | BT3

ANSWER KEY

11	**(a) The Louvain method [4 M]** The **Louvain method** (Blondel et al., 2008) is a fast, greedy, hierarchical algorithm that finds communities by **maximising modularity**, where modularity is $Q = \Sigma_c [L_c \div m - (d_c \div 2m)^2]$ (L_c = edges inside community c, d_c = sum of degrees of its nodes, m = total edges). At the start every node forms its own community. The algorithm then repeats two phases. **Phase 1 – Local moving of nodes.** The nodes are visited one by one. For each node i the algorithm calculates the **modularity gain ΔQ** that would result from removing i from its own community and placing it into the community of each of its neighbours: $\Delta Q = [(\Sigma_{in} + k_{i,in}) \div 2m - ((\Sigma_{tot} + k_i) \div 2m)^2] - [\Sigma_{in} \div 2m - (\Sigma_{tot} \div 2m)^2 - (k_i \div 2m)^2]$ where Σ_{in} is the total weight of edges inside the target community, Σ_{tot} is the sum of degrees of its nodes, k_i is the degree of node i and $k_{i,in}$ is the weight of links from i to nodes of that community. Node i is moved to the community with the largest **positive** gain; if no gain is positive it stays where it is. The sweep over all nodes is repeated until no move improves modularity (a local maximum). **Phase 2 – Community aggregation.** Each community found in Phase 1 is collapsed into a single **super-node**. The weight of the edge between two super-nodes is the sum of the weights of all edges between their members, and edges inside a community become a **self-loop**. This produces a much smaller weighted network.  *Fig. 1: The two phases of the Louvain method illustrated on the graph of Q.11* **Iteration.** Phase 1 and Phase 2 are repeated on the aggregated network until modularity no longer increases. Each pass gives one level of a hierarchy of communities, so the user can choose the resolution. **(b) Modularity of the two partitions [5 M]** Edges: 1–2, 1–3, 2–3, 2–4, 3–4, 4–5, 4–6, 5–6, so m = 8 and 2m = 16. Degrees: d(1) = 2, d(2) = 3, d(3) = 3, d(4) = 4, d(5) = 2, d(6) = 2 (sum = 16 = 2m). $Q = \Sigma_c [L_c \div m - (d_c \div 2m)^2]$ *P1: Q = 0.1250 + 0.1250 = 0.250 P2: Q = 0.0273 − 0.0664 + 0.0625 = 0.0234* **(c) Better partition [1 M]** **P1 is clearly the better community structure (Q = 0.25 against 0.023).** In P1 each community contains a complete triangle (three internal edges each) and only two edges (2–4 and 3–4) connect the two communities. In P2 the node pair {3, 4} contains only one internal edge but is separated from its triangle-mates, so P2 cuts through dense groups: its internal edge density is barely better than what a random network with the same degrees would give (Q close to 0). A check over all 203 possible partitions of the six nodes confirms that P1 gives the highest modularity, so it is also the partition that the Louvain method finds. (Q = 0.25 is below the usual 0.3 mark only because the network is very small and the extra edge 2–4 links the two triangles.)	10	CO2	BT3

Advantages	Limitations
Very fast, nearly linear in the number of edges; works on networks with millions of nodes	Result depends on the order in which nodes are processed (heuristic, not exact)
No need to specify the number of communities in advance	Resolution limit: very small communities may be merged
Produces a hierarchy of communities and handles weighted graphs	Communities are non-overlapping

Partition	Community	L_c	d_c	$L_c \div m$	$(d_c \div 16)^2$	Term
P1	{1, 2, 3}	3 (1–2, 1–3, 2–3)	2 + 3 + 3 = 8	0.3750	0.2500	0.1250
P1	{4, 5, 6}	3 (4–5, 4–6, 5–6)	4 + 2 + 2 = 8	0.3750	0.2500	0.1250
P2	{1, 2}	1 (1–2)	2 + 3 = 5	0.1250	0.0977	0.0273
P2	{3, 4}	1 (3–4)	3 + 4 = 7	0.1250	0.1914	−0.0664
P2	{5, 6}	1 (5–6)	2 + 2 = 4	0.1250	0.0625	0.0625

PART E: Answer Any 1 Following Questions. 10 M

Q.N.	Questions	M	CO	BT

ANSWER KEY

(a) Evolution of social networks [6 M]

A social network is not static: users join and leave, and new links are created and old ones vanish. **Network evolution** studies these changes over time. Analysis of many real networks (citation, email, online social and communication networks) shows several regular patterns, and network models explain the mechanisms behind them.

1. Densification [1.5 M]. As a network grows, the number of edges increases **faster than the number of nodes**.

Leskovec, Kleinberg and Faloutsos found that $e(t) \propto n(t)^a$ with an exponent $1 < a < 2$ (in the networks they studied about 1.1 to 1.7). Because $a = 1$ would mean constant average degree, $a > 1$ shows that the *average degree keeps increasing*: new nodes bring more links, and older nodes also form new links among themselves. Networks therefore become denser, not sparser, as they mature.

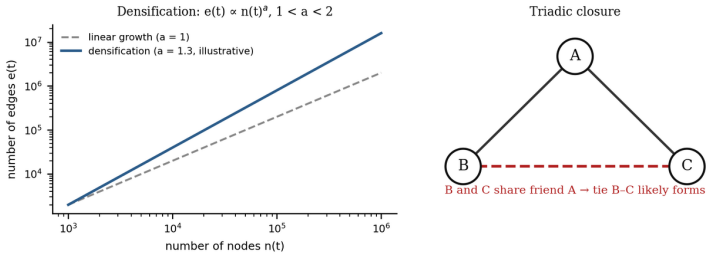


Fig. 1: Densification (left, illustrative) and triadic closure (right)

2. Shrinking diameter [1.5 M]. Contrary to intuition, the effective diameter (the distance within which most node pairs are connected) tends to **decrease or stabilise** as the network grows, because new links provide shortcuts between distant parts. This keeps the small-world property intact: even very large networks stay reachable in only a few hops.

3. Preferential attachment [1.5 M]. New nodes tend to connect to nodes that already have many links. In the Barabási–Albert model a new node links to an existing node i with probability proportional to its degree:

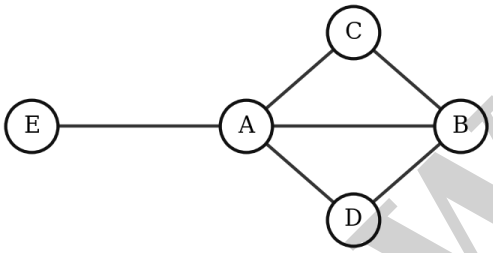
$P(i) = k_i \div \Sigma_j k_j$

This *rich-get-richer* process is the main explanation for the **power-law degree distribution** ($p(k) \propto k^{-3}$ in the basic model) and for the appearance of hubs. It also gives an advantage to older nodes, which have had more time to gather links.

4. Triadic closure [1.5 M]. If two users share a common friend, there is a strong tendency for them to become connected: an open triad (B–A–C) is "closed" by the new link B–C (see Fig. 1, right). The reasons are increased opportunity to meet, trust through the common friend and the social pressure to keep relationships balanced (Rapoport; Granovetter). Triadic closure **raises the clustering coefficient** and is the basis of friend-recommendation and link-prediction methods.

Analysis. The four mechanisms act together: preferential attachment builds hubs and a heavy-tailed degree distribution, triadic closure builds dense local clusters, densification raises the average degree, and the resulting shortcuts shrink the diameter. Homophily (similar users link), churn and platform features such as recommendation systems also shape evolution. Understanding these regularities allows analysts to forecast growth, detect anomalous changes (for example bot-driven bursts of links) and design recommendation policies.

(b) Preferential attachment for the new user F [4 M]



Network for Q.12 (edges: A–B, A–C, A–D, A–E, B–C, B–D)

(i) Degrees and attachment probabilities [2 M]. The network has 6 edges, so the sum of degrees is $\Sigma k = 2 \times 6 = 12$.

Node	Neighbours	Degree k	P(F links to node) = k ÷ 12	Decimal
A	B, C, D, E	4	4/12 = 1/3	0.333
B	A, C, D	3	3/12 = 1/4	0.250
C	A, B	2	2/12 = 1/6	0.167
D	A, B	2	2/12 = 1/6	0.167
E	A	1	1/12	0.083
Total		12	1	1.000

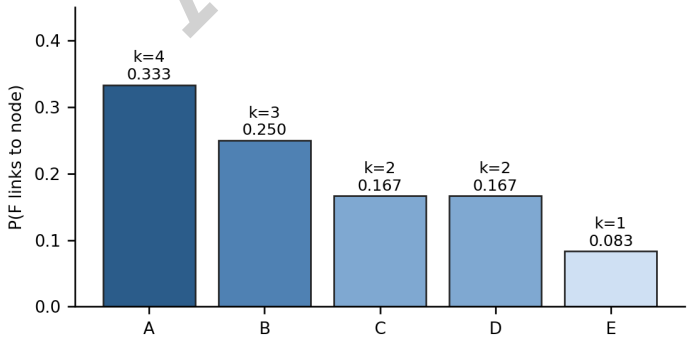


Fig. 2: Probability that F attaches to each existing node

(ii) Effect on the degree distribution [2 M]. Because the probability of receiving the new link is proportional to the current degree, node A is four times as likely to be chosen as node E. Every new link therefore raises the chance that the same node will be chosen again – a *rich-get-richer* effect. Repeated over many arrivals, this produces a **heavy-tailed (power-law) degree distribution**: a few nodes grow into very large hubs, whereas the majority of nodes keep a very small degree. The inequality in the number of connections keeps increasing, the mean degree is not representative, and the influence of a few hub users becomes disproportionately large. The network also becomes vulnerable to the removal of hubs but resilient to the removal of random nodes.

12		10	CO2	BT4

ANSWER KEY

(a) Relationship mining and tie strength [4 M]

Relationship mining. Relationship mining is the process of **discovering and characterising the relationships between users (or other entities) from interaction and co-occurrence data** – for example, inferring close friends, colleagues, family or collaborators from message counts, comments, tags, co-attendance and shared activities. It also uncovers *implicit* or hidden relationships that users have never declared explicitly, and it is the basis of interaction networks in which edge weights represent the intensity of interaction.

Tie strength. Tie strength expresses how strong a relationship is. Granovetter (1973) described it as a combination of the amount of time, emotional intensity, intimacy and reciprocal services that characterise the tie. On social media it is estimated from observable indicators such as **frequency and recency of interaction, reciprocity, duration of the relationship, the number of mutual friends and the type of interaction** (a private message signals a stronger tie than a like). Studies such as Gilbert and Karahalios (2009) showed that tie strength can be predicted from social media data.

Basis	Strong ties	Weak ties
Nature	Close friends, family, close colleagues	Acquaintances, distant contacts, occasional collaborators
Interaction	Frequent, reciprocal, emotionally intense	Infrequent, casual, low intensity
Network position	Lie inside dense clusters; friends share many mutual contacts	Often act as **bridges** between different clusters
Information	Mostly redundant, because the same circle knows the same things	Novel, non-redundant information and opportunities – "the strength of weak ties"
Main functions	Emotional support, trust, influence, group cohesion	Spread of information and innovations across groups; job leads; network reach

(b) Analysis of the interaction network

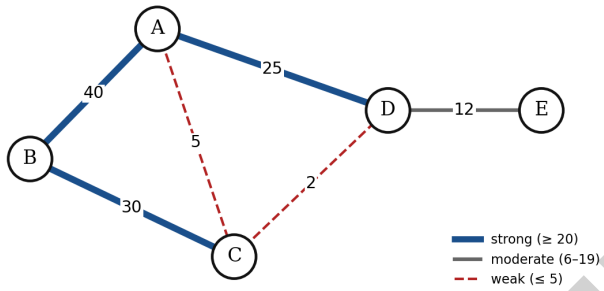


Fig. 1: Interaction network (edge label = interactions per month)

(i) Classification of ties [2 M]. Rule: interactions ≥ 20 are strong; 6–19 are moderate; ≤ 5 are weak.

Tie	Interactions per month	Class	Reason
A–B	40	**Strong**	40 ≥ 20
A–C	5	**Weak**	5 ≤ 5
A–D	25	**Strong**	25 ≥ 20
B–C	30	**Strong**	30 ≥ 20
C–D	2	**Weak**	2 ≤ 5
D–E	12	**Moderate**	6 ≤ 12 ≤ 19

(ii) Weighted degree (interaction strength) [2 M]. The weighted degree (or strength) of a user is the **sum of the weights of all ties of that user**.

User	Ties and weights	Weighted degree	Rank
A	A–B 40 + A–C 5 + A–D 25	**70**	1 (tie)
B	A–B 40 + B–C 30	**70**	1 (tie)
D	A–D 25 + C–D 2 + D–E 12	39	3
C	A–C 5 + B–C 30 + C–D 2	37	4
E	D–E 12	12	5

The most interactive users are **A and B (weighted degree 70 each)**. They are linked by the strongest tie in the network (40) and, together with C, form a closely interacting core (A–B and B–C are strong). E is the least interactive user (12). Note that the ordinary (unweighted) degree of A is 3, that of B is 2 and that of D is 3, so weights give a different and more realistic picture of activity than counting ties alone.

(iii) The user whose removal disconnects the network [2 M]. The network consists of the edges A–B, A–C, A–D, B–C, C–D and D–E. Testing the removal of each user:

User removed	Remaining edges	Remaining users connected?
A	B–C, C–D, D–E	Yes (B–C–D–E)
B	A–C, A–D, C–D, D–E	Yes
C	A–B, A–D, D–E	Yes (B–A–D–E)
D	A–B, A–C, B–C (E has no edge left)	**No – E is isolated**
E	A–B, A–C, A–D, B–C, C–D	Yes

Only the removal of **user D** disconnects the network: D is a **cut vertex (articulation point)** and the moderate tie D–E is a bridge, because E is connected to the rest of the network only through D. D is therefore a critical broker – if D stops interacting, E is cut off from A, B and C.

Additional insight (strong triadic closure). User A has two strong ties (to B and to D), but B and D are not connected at all. According to the strong triadic closure property, a tie between B and D would be expected; it is therefore a good

13 | 10 | CO2 | BT4

ANSWER KEY