Network analysis

Note on data. This problem set uses **synthetic** (simulated) network data generated in the provided Python tutorial script. The goal is to practice network workflows (construction, centrality, communities, and GNN-style node classification)—not to draw substantive inferences about real-world actors.

Conceptual Questions

Please write three to ten sentence explanations for each of the following questions. **You are only required to answer ONE of the two questions below.**

1. Networks require representation choices (directed vs. undirected, weighted vs. unweighted, missing edges vs. absent edges, handling isolates). Choose **two** representation choices and explain how each could change a downstream result in this lab (e.g., centrality rankings, community detection, or node classification). Use one concrete example for each.
2. Graph neural networks (GNNs) combine node features and network structure via message passing. Explain, conceptually, why using the graph can improve prediction beyond a feature-only model. Then identify one way a GNN can *fail* or mislead in social science settings (e.g., homophily-driven overconfidence, label leakage, missing data, boundary specification).

Applied Exercises

Use the code in the week's code tutorial and the lecture slides to answer the following questions.

3. **Graph construction + centrality.** Using the provided Python script:
   - Generate the synthetic stochastic block model (SBM) network and save (or print) a basic summary: number of nodes, number of edges, and density.
   - Create an edge list (`u`, `v`) and reconstruct the graph from the edge list.
   - Compute three centrality measures:
     (a) degree (or degree centrality),
     (b) approximate betweenness centrality (using sampling),
     (c) eigenvector centrality.
   - Create **one** figure that shows the distribution of each centrality measure (three histograms is fine).
   - Report the top 10 nodes under each measure and briefly discuss (4–8 sentences): do the rankings agree? What kind of "importance" does each metric capture in this synthetic network?
4. **Community detection + evaluation.** Using the same synthetic network:
   - Run Louvain community detection and report:
     (a) the number of detected communities, and

(b) the sizes of the communities (a small table is fine).

- Compute the Adjusted Rand Index (ARI) comparing Louvain communities to the known SBM ground-truth labels.
- Create **one** visualization that communicates the community structure result (e.g., bar plot of community sizes).
- In 5–8 sentences, interpret what the ARI means here and give one reason why community detection might split or merge true blocks (even when the data are generated from an SBM).

5. **Node classification: features-only baseline vs. GCN-style model.** Using the node features and ground-truth labels in the tutorial script:
   - Fit a features-only baseline model (logistic regression) and report validation and test accuracy.
   - Train the 2-layer GCN-style model from the tutorial and report validation and test accuracy.
   - Create a confusion table (or confusion matrix) for the test set for the GCN model (a table is sufficient).
   - In 6–10 sentences, compare the baseline and GCN results. Your discussion must include:
     (a) why the GCN can outperform the baseline in this setting,
     (b) one reason the baseline might be competitive (or even better) in some settings, and
     (c) one caution about interpreting "high accuracy" as scientific validity in social network prediction tasks.
6. **Challenge Question (Optional — if you finish early):** Run a small sensitivity analysis that varies either **network homophily** or **feature noise**, and compare community detection and prediction performance.
   - Choose **one** knob to vary:
     (a) increase/decrease the off-diagonal probabilities in the SBM matrix P (more vs. less between-community mixing), *or*
     (b) increase/decrease the feature noise standard deviation (`noise_sd`).
   - Run **three** settings (low / medium / high) and record:
     (a) Louvain ARI,
     (b) baseline logistic regression test accuracy,
     (c) GCN test accuracy.
   - Present a small summary table (and optionally a plot) and interpret the pattern (5–10 sentences). In your interpretation, explain when network structure helps most and when it helps least.