**Project Instructions for Social Network Analysis Using Gephi**

### **General Instructions**

1. **Group Formation:**
   * Form groups of up to 4 members.
   * Assign roles within the group (e.g., data preparation, analysis, visualization, documentation).
2. **Dataset Selection:**
   * Each group must choose one dataset from the provided list.
   * Review the dataset documentation (links provided) to understand its structure and context.
3. **Tasks:**

a. **In-Degree and Out-Degree Calculation:**

* + For each node, calculate the in-degree (number of edges pointing to the node) and out-degree (number of edges originating from the node).

b. **Node Ranking:**

* + Rank nodes based on their activity level. Use metrics such as the sum of in-degree and out-degree to define activity.

c. **Top and Bottom Nodes Selection:**

* + Identify the top 20 most active nodes and the 20 least active nodes.
  + Create a separate dataset containing only these 40 nodes and their connections.

d. **Network Visualization:**

* + Import the working dataset into Gephi.
  + Create a visual representation of the network using Gephi’s tools.
  + Experiment with layout algorithms (e.g., ForceAtlas2, Yifan Hu).
  + Use node and edge properties (e.g., color, size, width) to highlight key features.

e. **Interpretation:**

* + Analyze the visualized network.
  + Discuss the structural differences between active and inactive nodes, subgroups, and their positions in the network.

1. **Submission format:**
   * Prepare a presentation (7-8 slides) summarizing:
     + Key metrics (in-degree, out-degree, etc.).
     + Graph visualizations from Gephi.
     + Interpretations of the results.
   * Include screenshots of Gephi visualizations in the slides.
   * Submit the slides in PDF format.
   * Ensure all group member names and roles are included in the submission.
   * Due date: **January 24, 23:59 on C@mpus Moodle**.
2. **Evaluation Criteria**
   * **Methodology (30%)**
     + Clarity and correctness of steps.
     + Calculation and ranking process for nodes.
   * **Visualization (40%)**
     + Quality and clarity of Gephi visualizations.
     + Effective use of Gephi features (e.g., layouts, colors, sizes).
   * **Interpretation (30%)**
     + Depth of analysis and insights.
     + Dataset-specific questions addressed comprehensively.

### **Datasets**

#### **1. Wikipedia Voting Network**

* URL: <https://snap.stanford.edu/data/wiki-Vote.html>
* **Questions:**
  + How many nodes and edges are present in the network?
  + What is the distribution of in-degree and out-degree across the nodes?
  + Identify nodes with the highest in-degree and out-degree. What roles might they play in the voting process?
  + Are there any bridge nodes that connect otherwise disconnected components?
* **Conclusion Question:**
  + What does the structure of this network reveal about influence and participation in voting dynamics?

#### **2. Math Overflow Network**

* URL: <https://snap.stanford.edu/data/sx-mathoverflow.html>
* **Questions:**
  + What is the total number of nodes and edges?
  + How does the network exhibit multiplexity (e.g., multiple interactions between nodes)?
  + Identify the central nodes using degree centrality and betweenness centrality. What roles do they play?
  + Are there clusters or communities in the network? What might they represent?
* **Conclusion Question:**
  + How does this network structure reflect patterns of knowledge exchange in a Q&A platform?

#### **3. Email EU Core Network**

* URL: <https://snap.stanford.edu/data/email-Eu-core-temporal.html>
* **Questions:**
  + What is the size of the network (nodes and edges)?
  + What is the average in-degree and out-degree of the nodes?
  + Identify the most central nodes based on betweenness centrality. What is their significance in communication flow?
  + Are there isolated nodes or small components? What do they signify?
* **Conclusion Question:**
  + How does the temporal aspect of the network affect the flow of communication?

#### **4. College Messaging Network**

* URL: <https://snap.stanford.edu/data/CollegeMsg.html>
* **Questions:**
  + How many nodes and edges exist in the network?
  + What is the distribution of activity levels (in-degree and out-degree)?
  + Identify any bridge nodes that connect major communities.
  + Are there densely connected clusters? What might they indicate?
* **Conclusion Question:**
  + What does this network reveal about social dynamics and communication in a college setting?

#### **5. Bitcoin OTC Trust Network**

* URL: <https://snap.stanford.edu/data/soc-sign-bitcoin-otc.html>
* **Questions:**
  + What is the size of the network (nodes and edges)?
  + How do in-degree and out-degree distributions vary among nodes?
  + Which nodes are the most trusted based on centrality metrics?
  + Are there communities or clusters? What might these represent in the context of trust?
* **Conclusion Question:**
  + How does the network structure reflect patterns of trust and risk in financial transactions?

#### **6. Enron Email Network**

* URL: <https://snap.stanford.edu/data/email-Enron.html>
* **Questions:**
  + How many nodes and edges are in the dataset?
  + What is the distribution of in-degree and out-degree?
  + Identify key individuals (nodes) based on centrality metrics.
  + Are there isolated components or key bridges? What do they signify?
* **Conclusion Question:**
  + What does this network reveal about organizational communication and hierarchy?

#### **7. Reddit Hyperlink Network**

* URL: <https://snap.stanford.edu/data/soc-RedditHyperlinks.html>
* **Questions:**
  + What is the size of the network in terms of nodes and edges?
  + Which subreddits are the most central (based on degree or betweenness centrality)?
  + How does multiplexity appear in this dataset?
  + Are there clear clusters or communities? What do they represent?
* **Conclusion Question:**
  + How does the structure of this network reflect content sharing and interaction patterns across subreddits?