# GNotes - not formatted

💡 definition of centrality measure and definingeach centrality measure along with how we incorporated/used each centrality measure for TfL network and what does that mean for the whole network

## Centrality Measure

Centrality measures are metrics used in network analysis to identify the most important nodes within a graph. These measures help determine which nodes are critical for the network's structure and functioning. Various centrality measures capture different aspects of a node's importance.

## Definitions and Mathematical Formulas

1. **Degree Centrality**:
   - **Definition**: Degree centrality measures the number of direct connections a node has. It is a simple count of how many edges are incident to a node.
   - **Formula**:CD(v)=deg(v)

     CD(v)=deg(v)C_D(v) = \text{deg}(v)

     where deg(v)\text{deg}(v)deg(v) is the degree of node vvv.
   - **Manual Calculation**:

     ```
     degree_centrality = {node: len(list(G.neighbors(node)))
     for node in G.nodes()}
     ```

     This code calculates the degree centrality for each node by counting its neighbors.
   - **Interpretation for TfL Network**: Nodes with high degree centrality are major stations with many direct connections to other stations. They are crucial for direct travel within the network.
2. **Betweenness Centrality**:

- **Definition**: Betweenness centrality measures the extent to which a node lies on the shortest paths between other nodes. It quantifies the control a node has over information flow in the network.
- **Formula**:CB(v)=s≠v≠t∑σstσst(v)

  CB(v)=∑s≠v≠tσst(v)σstC_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}

  where σst\sigma_{st}σst is the total number of shortest paths from node sss to node ttt, and σst(v)\sigma_{st}(v)σst(v) is the number of those paths that pass through vvv.
- **Manual Calculation**:

  ```
  betweenness_centrality = nx.betweenness_centrality(G)
  ```

  NetworkX calculates this using an efficient algorithm, but conceptually, it involves finding all shortest paths and counting how many times each node appears on these paths.
- **Interpretation for TfL Network**: Nodes with high betweenness centrality are key transfer points, facilitating travel between different parts of the network.

3. **Closeness Centrality**:
   - **Definition**: Closeness centrality measures how close a node is to all other nodes in the network. It is the reciprocal of the average shortest path distance to all other nodes.
   - **Formula**:CC(v)=∑u≠vd(v,u)1

     CC(v)=1∑u≠vd(v,u)C_C(v) = \frac{1}{\sum_{u \neq v} d(v, u)}

     where d(v,u)d(v, u)d(v,u) is the shortest path distance between nodes vvv and uuu.
   - **Manual Calculation**:

     ```
     closeness_centrality = nx.closeness_centrality(G)
     ```

     This involves calculating the shortest path from each node to every other node and taking the reciprocal of the average distance.
   - **Interpretation for TfL Network**: Nodes with high closeness centrality are centrally located, making them efficient for travel to other parts of the network.

4. **Eigenvector Centrality**:

   - **Definition**: Eigenvector centrality measures a node's influence based on the influence of its neighbors. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question.
   - **Formula**:CE(v)=λ1u∈N(v)∑CE(u)

     CE(v)=1λ∑u∈N(v)CE(u)C_E(v) = \frac{1}{\lambda} \sum_{u \in N(v)} C_E(u)

     where λ\lambdaλ is a constant, N(v)N(v)N(v) are the neighbors of vvv, and CE(u)C_E(u)CE(u) is the eigenvector centrality of node uuu.
   - **Manual Calculation**:

     ```
     eigenvector_centrality = nx.eigenvector_centrality(G)
     ```

     This is calculated using an iterative algorithm until the centrality values converge.
   - **Interpretation for TfL Network**: Nodes with high eigenvector centrality are well-connected to other well-connected nodes, indicating influential stations in terms of network connectivity.

5. **PageRank**:

   - **Definition**: PageRank is a variant of eigenvector centrality originally designed for ranking web pages. It measures the importance of a node based on the structure of incoming links and the importance of the source nodes.
   - **Formula**:PR(v)=N1−d+du∈M(v)∑L(u)PR(u)

     PR(v)=1−dN+d∑u∈M(v)PR(u)L(u)PR(v) = \frac{1-d}{N} + d \sum_{u \in M(v)} \frac{PR(u)}{L(u)}

     where ddd is the damping factor, NNN is the total number of nodes, M(v)M(v)M(v) are the nodes linking to vvv, and L(u)L(u)L(u) is the number of outbound links from node uuu.
   - **Manual Calculation**:

     ```
     pagerank = nx.pagerank(G)
     ```

This uses an iterative algorithm similar to eigenvector centrality but includes the damping factor to ensure convergence.

- **Interpretation for TfL Network**: Nodes with high PageRank are important not just because of their connections but also due to the quality and influence of the nodes they connect to.

## Combining Centrality Measures

To identify nodes with high centrality across multiple measures, we:

- Calculate each centrality measure.
- Extract the top N nodes for each measure.
- Combine these lists and count occurrences of each node.

Mathematically, we use:

$S(v) = \sum_{i=1}^{5} \delta(v \in T(C_i))$

$$S(v) = \sum_{i=1}^{5} \delta(v \in T(C_i))$$

where $T(C_i)$ is the set of top nodes for centrality measure $C_i$, and $\delta$ is the indicator function.

## Insights and Importance of These Nodes

Nodes identified as central across multiple measures are crucial for the network's resilience, efficiency, and robustness:

- **Kings Cross St Pancras**: A major interchange, central for connectivity.
- **Baker Street**: Multiple connections make it a key hub.
- **Oxford Circus**: High traffic and central location.
- **Green Park**: Important for transfers and network flow.
- **Stratford**: Growing significance with new developments.
- **Waterloo**: One of the busiest and most connected stations.
- **Westminster**: Central for politics and tourism.
- **Bond Street**: Key for commercial and business districts.

These nodes ensure smooth operation and connectivity within the TfL network, making them focal points for network management, development, and emergency planning.

---

> 💡 Nodes with the highest centrality across multiple measures:
> ['kingscrossstpancras', 'bakerstreet', 'oxfordcircus', 'greenpark', 'stratford', 'waterloo', 'westminster', 'bondstreet']
>
> calculation of combining all the Nodes with the highest centrality across multiple measures

## Explanation of the Code and Concept

1. **Centrality Measures**:
   Centrality measures help identify the most important nodes in a network. Each measure captures different aspects of a node's importance:
   - **Degree Centrality**: Number of connections a node has.
   - **Betweenness Centrality**: Frequency at which a node lies on the shortest path between other nodes.
   - **Closeness Centrality**: Inverse of the sum of distances from a node to all other nodes.
   - **Eigenvector Centrality**: Influence of a node in terms of the centrality of its neighbors.
   - **PageRank**: A variant of Eigenvector Centrality, used to rank nodes based on their connections.
2. **Calculation of Top Nodes**:
   - We compute each centrality measure using NetworkX functions.
   - For each measure, we extract the top N nodes. This gives us the most important nodes according to each centrality measure.
3. **Combining Top Nodes**:
   - We create a list that combines the top nodes from all centrality measures.
   - Using a count of occurrences, we identify nodes that appear multiple times across different centrality measures.

   - Nodes that appear frequently in these lists are considered central across multiple dimensions of importance.

4. **Mathematical Formula for Combined Centrality**:
   Let's denote the set of top nodes for each centrality measure $C_i$ as $T(C_i)$. For example:

   - $T(C_1)$ for Degree Centrality
   - $T(C_2)$ for Betweenness Centrality
   - $T(C_3)$ for Closeness Centrality
   - $T(C_4)$ for Eigenvector Centrality
   - $T(C_5)$ for PageRank

   If we denote $T(C)$ as the combined top nodes set, then:
   $$
   T(C) = \bigcup_{i=1}^{5} T(C_i)
   $$

   The importance score for each node $v$ in the combined set is given by the count of its occurrences:
   $$
   S(v) = \sum_{i=1}^{5} \delta(v \in T(C_i))
   $$
   where $\delta$ is the indicator function, which is 1 if $v$ is in $T(C_i)$, and 0 otherwise.

   Nodes with the highest scores $S(v)$ are those appearing most frequently in the top lists of centrality measures.

## Insights and Importance of the Output

- **Key Nodes**:
  - **Kings Cross St Pancras**: A major interchange in the network, connecting multiple lines.
  - **Baker Street**: A significant hub with multiple connections.
  - **Oxford Circus**: A central station in the network with high traffic.

- **Green Park**: An important transfer point for passengers.
- **Stratford**: A growing hub, especially after recent developments.
- **Waterloo**: One of the busiest stations with multiple connections.
- **Westminster**: Central for political and tourist destinations.
- **Bond Street**: A key station for shopping and business districts.

- **Importance of These Nodes**:
  - These nodes are crucial for the efficient functioning of the network. They connect multiple lines, ensuring that passengers can transfer between different parts of the network smoothly.
  - High centrality nodes often indicate stations with high passenger volumes and significant traffic flow, making them critical for network resilience and robustness.
  - Understanding these nodes helps in planning and managing network operations, especially during peak times or disruptions.

## Why This Analysis is Essential

1. **Network Robustness**:
   Identifying central nodes helps in understanding points of vulnerability. Ensuring these nodes are well-maintained and secure can prevent major disruptions.
2. **Resource Allocation**:
   High centrality nodes may require more resources, such as staff and security, due to their critical role in the network.
3. **Planning and Development**:
   Insights from centrality measures can guide future expansions or modifications to the network. For instance, adding new connections to less central areas can help in balancing the network load.
4. **Emergency Management**:
   Knowing key nodes helps in effective emergency response planning, ensuring swift evacuation and management of high-traffic areas during crises.

## Conclusion

This analysis leverages multiple centrality measures to provide a comprehensive view of the most important nodes in the network. By combining these measures, we identify nodes that are consistently central, offering valuable insights for network management, planning, and operations. The output nodes, such as Kings Cross St Pancras and Baker Street, play a pivotal role in the network's efficiency and resilience, making them focal points for strategic planning and resource allocation.