Intro to Graphs

Vi Ly

3 Apr 2024

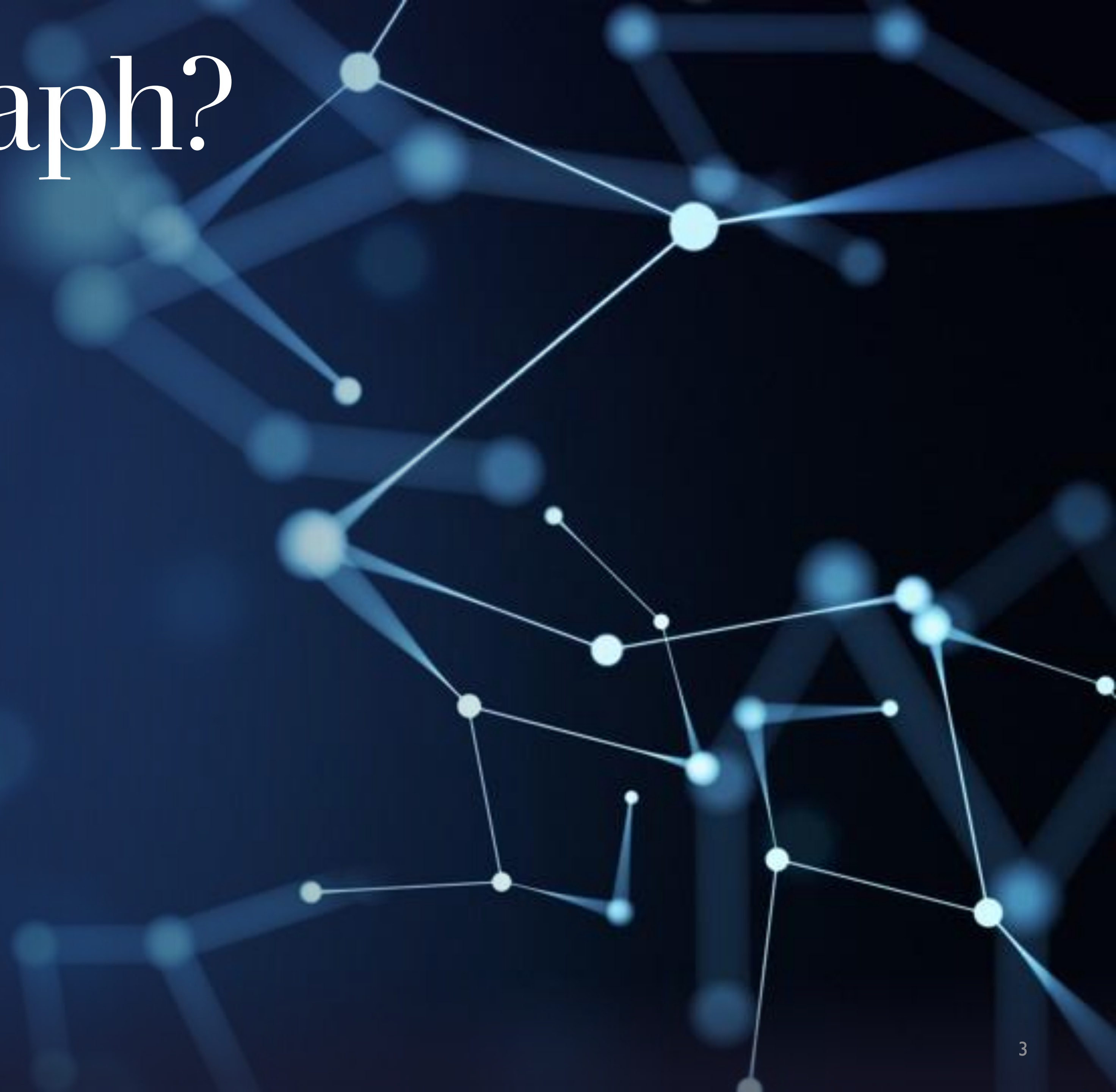
Agenda

- What is a Graph?
- Graph Representation
- Node Importance
- Additional Terminology
- Graph Coloring
- Use Case – Identifying Fraud Rings
- Jupyter Notebook Hands On

What is a Graph?

- Structure consisting of
  - Vertices (aka Nodes)
    - Entities
  - Edges (aka Links)
    - Relationships between nodes

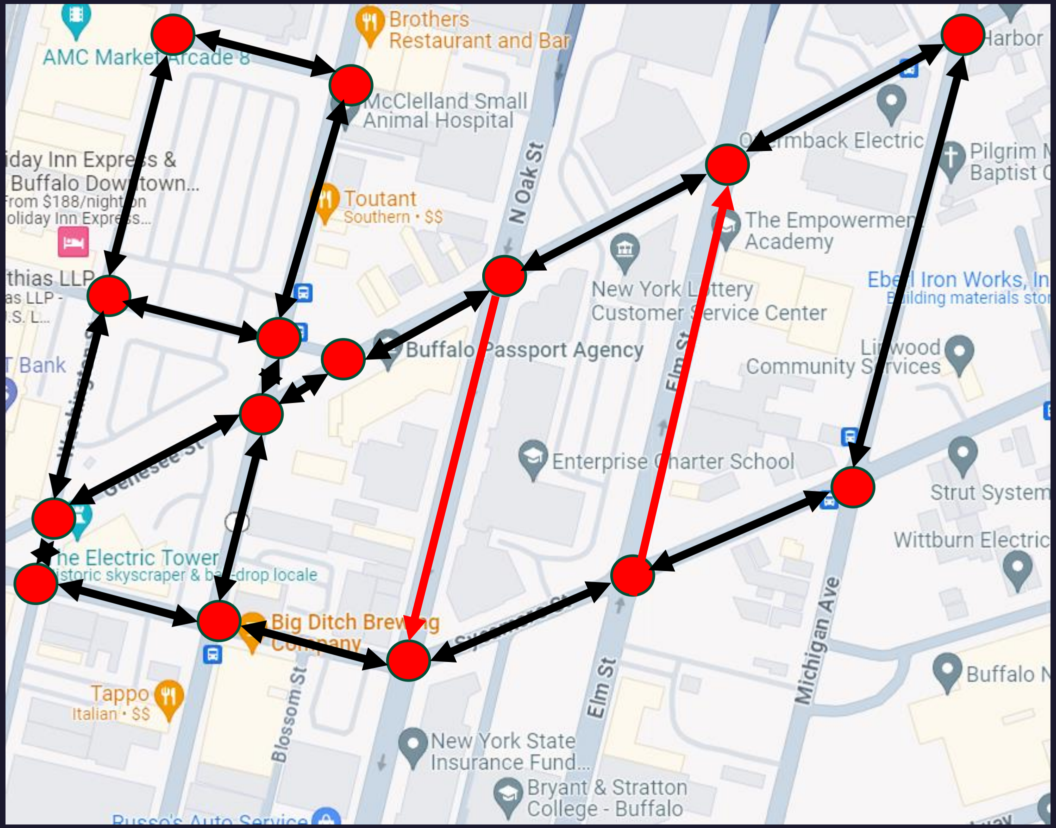
- Graphs can have only nodes and no edges
  - Inverse does not hold true



What is a Graph?

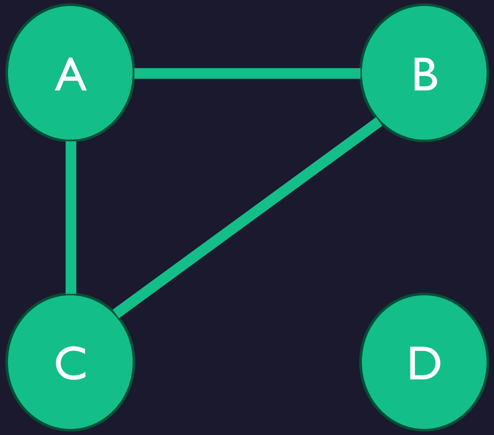
- Undirected vs Directed Graphs
- Weighted vs Unweighted Edges
  - Weights represent strength of relationship or cost of travel
  - Can have multiple sets of weights
    - Example – GPS
      - Distance weights
      - Travel Time weights
      - Cost (Gas & Tolls) weights
- Graph Use Cases
  - Social Networks
  - Causal Inference
  - DAGs
  - Trees
  - GPS

Directed Graph Example

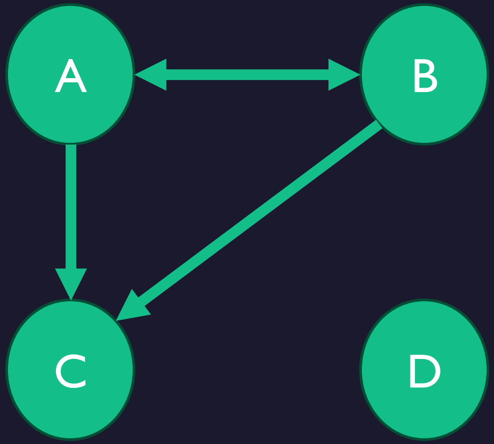


Graph Representation

- Adjacency Matrix
    - N x N Matrix where N is the number of nodes
    - Undirected Graphs – Matrix is symmetric
    - Directed Graphs – Matrix is not symmetric
    - Unweighted Graphs – 1 if edge exists
    - Weighted Graphs – Non-zero value for edge
- Not space efficient
- Matrix increases in dimension with new nodes



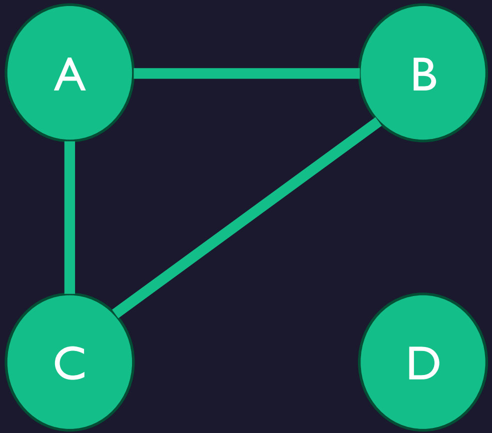
	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	1	1	0	0
D	0	0	0	0

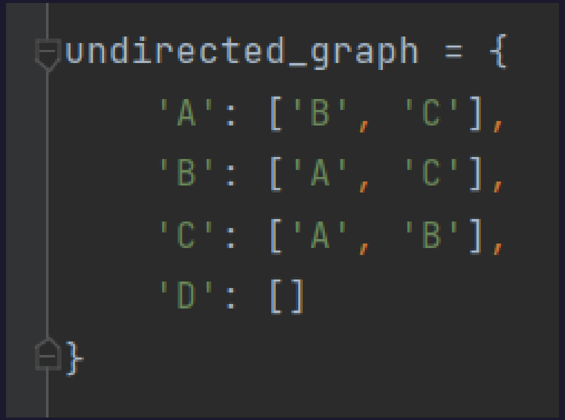


	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	0	0	0	0
D	0	0	0	0

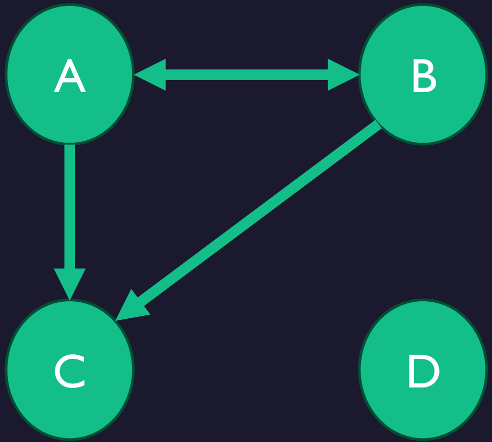
Graph Representation

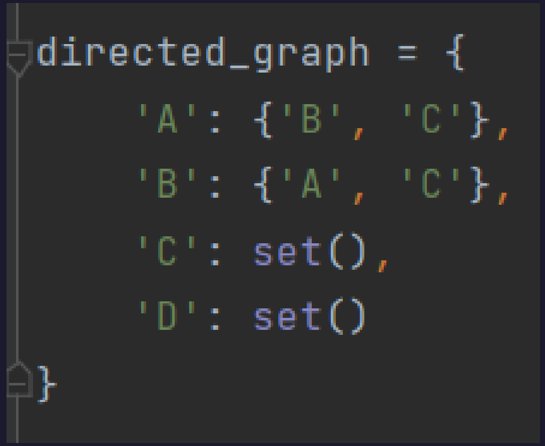
- Adjacency List (In Python)
  - Dict
    - Keys: Nodes
    - Values: Collection of Neighbors
- Most common implementation
- More space efficient compared to matrix





```
undirected_graph = {
        'A': ['B', 'C'],
        'B': ['A', 'C'],
        'C': ['A', 'B'],
        'D': []
}
```





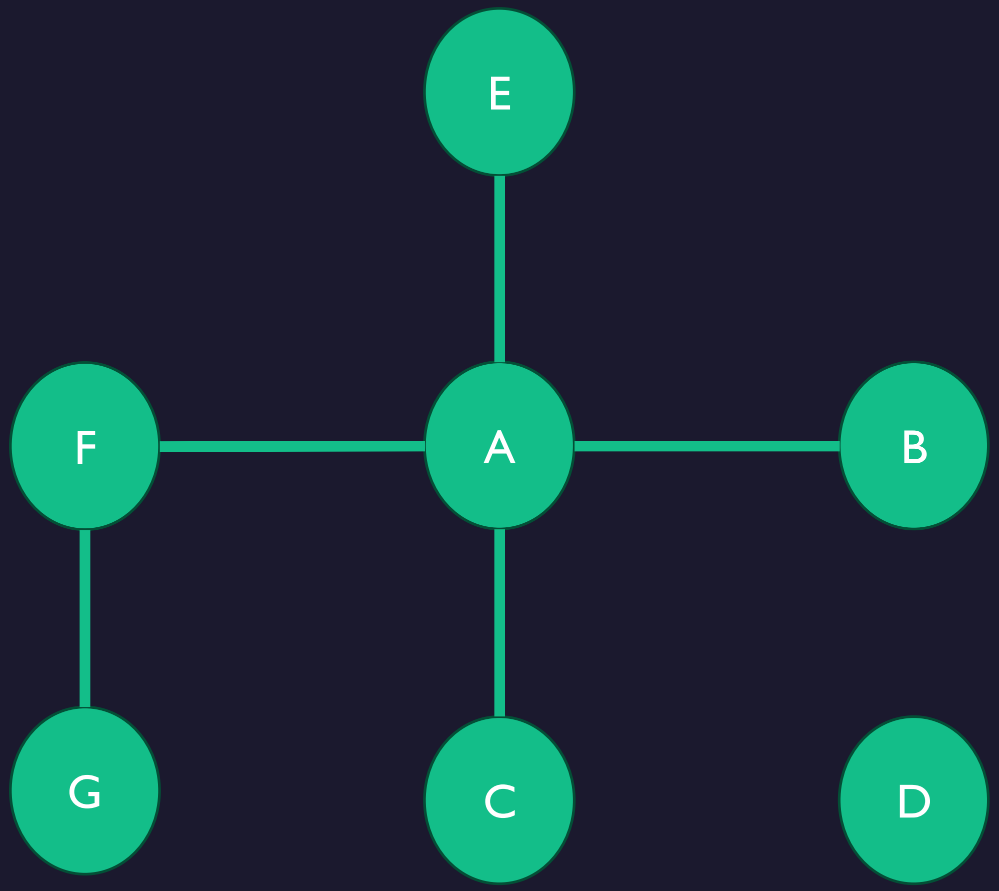
```
directed_graph = {
        'A': {'B', 'C'},
        'B': {'A', 'C'},
        'C': set(),
        'D': set()
}
```

Node Importance

- Identify “influencers”
- Many metrics for node importance
- 3 Common metrics
  - Degree Centrality
  - Betweenness Centrality
  - Eigenvector Centrality

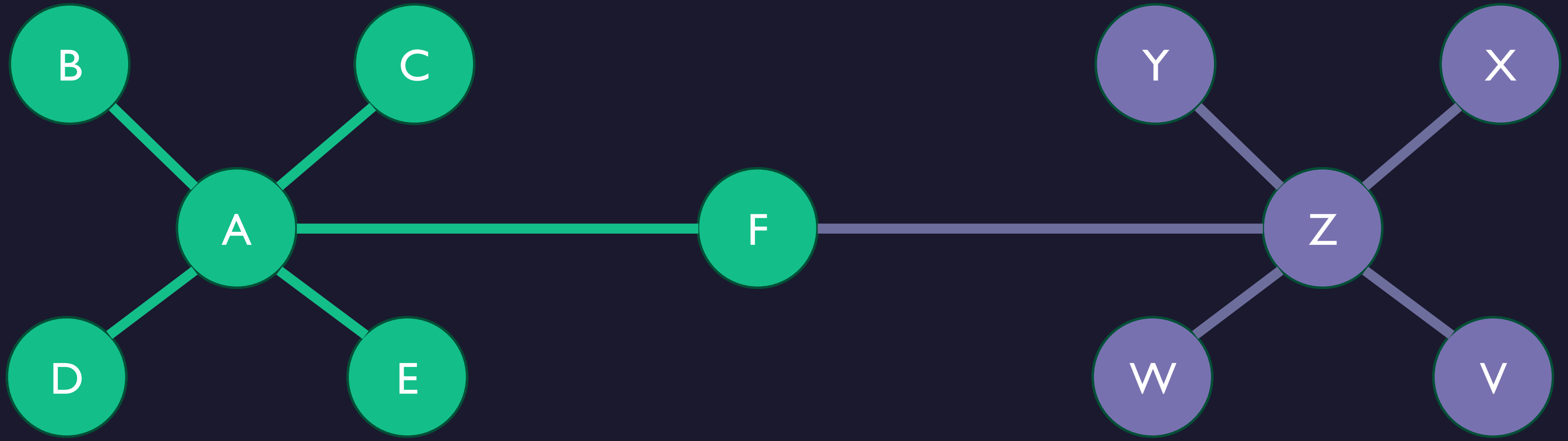
Degree Centrality

- # of edges connected to a node
- Node A has highest degree centrality
- Node F has next highest degree centrality
- Node D has lowest degree centrality
- Remaining nodes have the same degree centrality



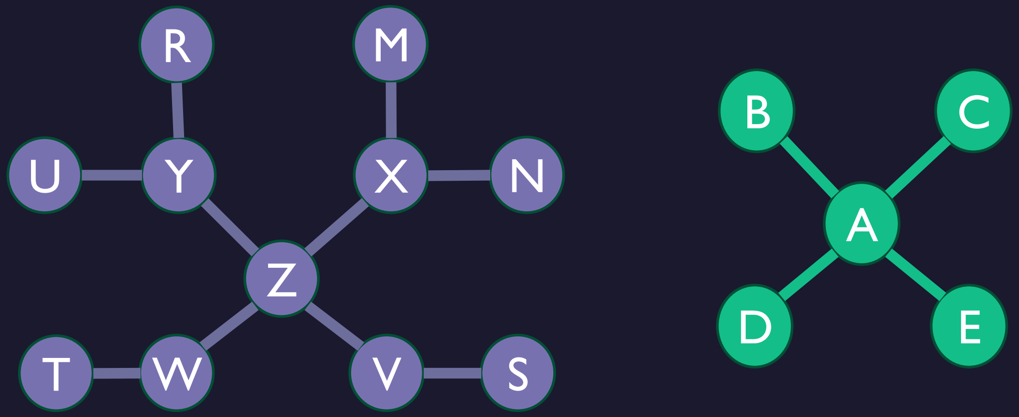
Betweenness Centrality

- How often is a node in the path between 2 other nodes
- # of shortest paths containing node
- Nodes A & Z have highest degree centrality
  - Node F has low degree centrality
  - Can we use a different metric to capture the importance of F?
- Very compute intensive – not feasible on large graphs without sampling



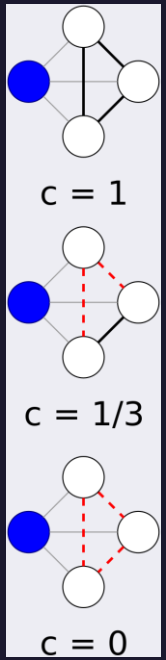
Eigenvector Centrality

- Nodes connected to other influential nodes
- Nodes A & Z have same degree centrality
- Node Z has highest eigenvector centrality



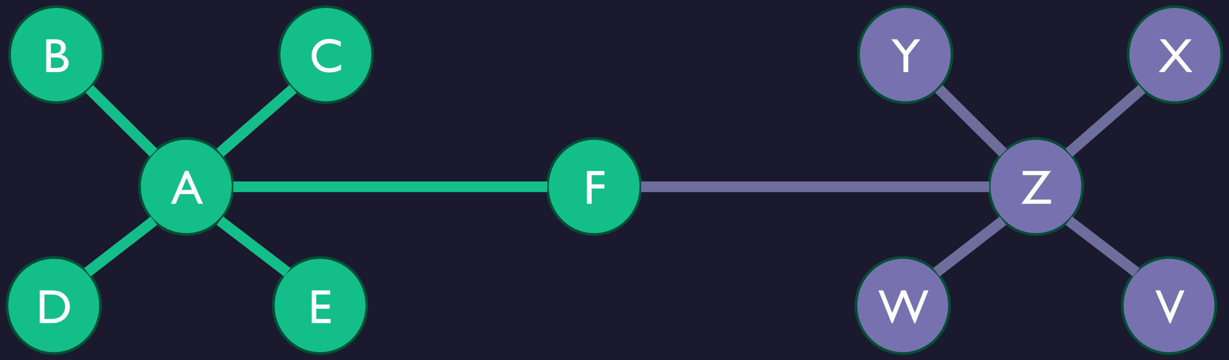
Clustering Coefficient

- For a given node, how connected are its neighbors?



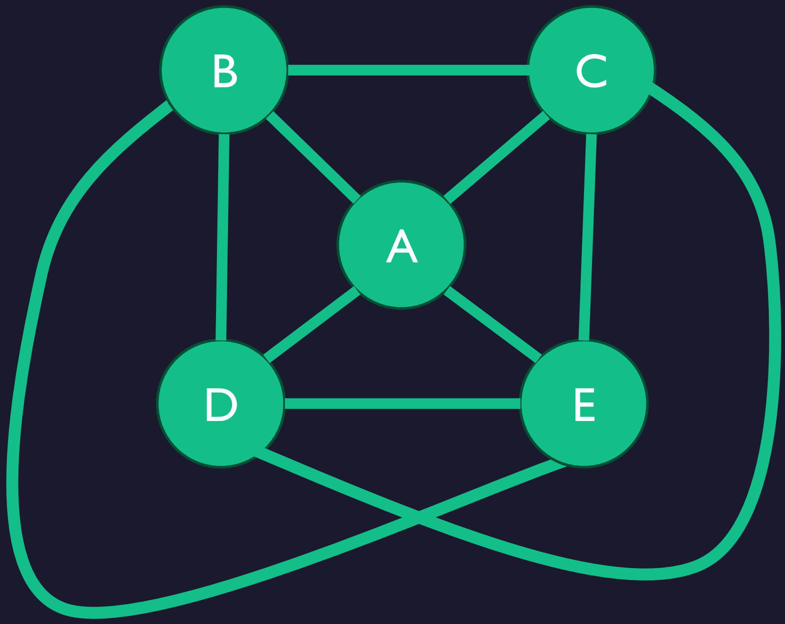
Additional Terminology

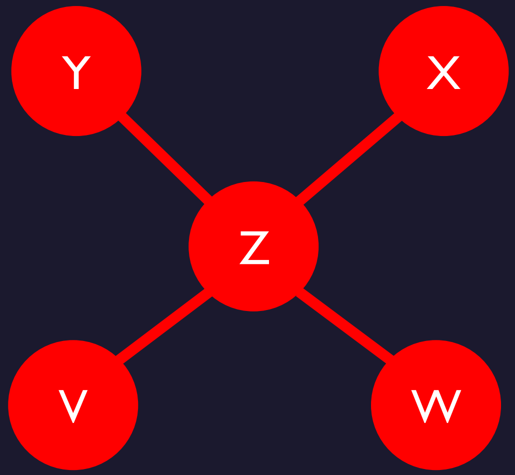
- Connected Graph
  - All nodes can be reached by every other node
    - Example is a connected graph
- Connected Component
  - Subgraph (subset of nodes and edges) which is connected
  - Connected Graph – 1 Connected Component
- Cut Vertex
  - Node whose removal results in additional connected component
  - Node F is a cut vertex.
    - Remove F and there will now be 2 connected components (green & purple)
  - Nodes A & Z are also cut vertices



Additional Terminology

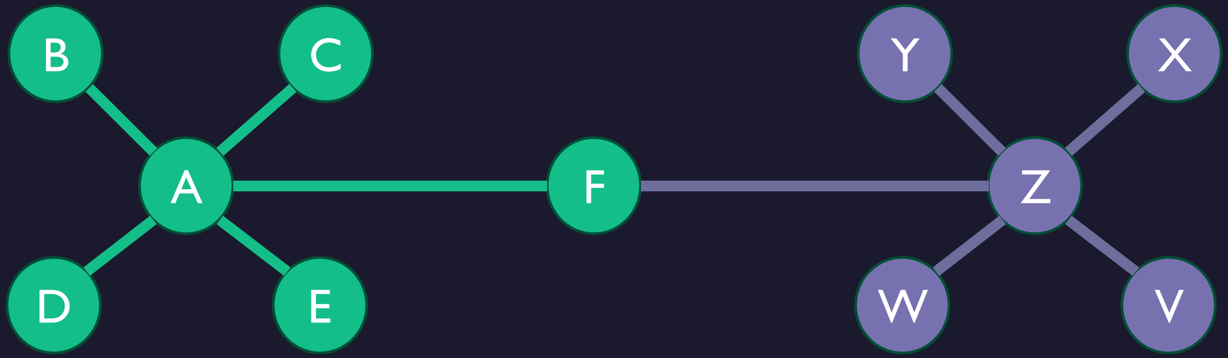
- Clique
  - Subset of nodes where each node is adjacent to every other node
    - Green subset is a clique
    - Red subset is not a clique
- Complete Graph
  - If the entire graph is a clique, it is a complete graph





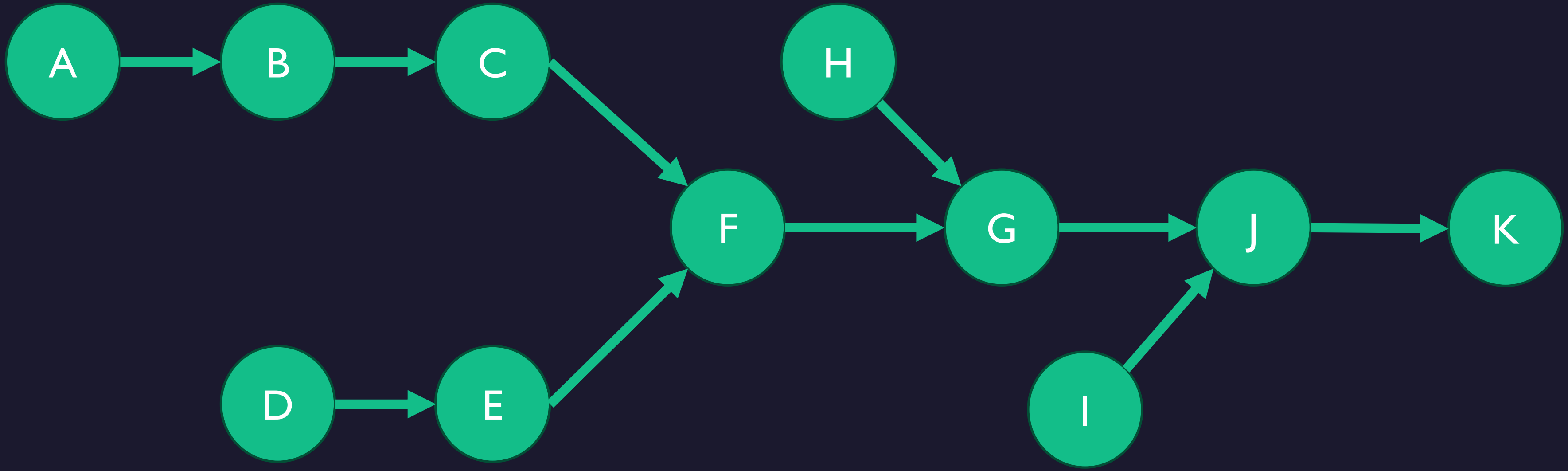
Community Detection

- Connected Graph
- 1 Connected Component
- 2 Communities
- How to identify communities
    - Many algorithms
    - Louvain Community Detection



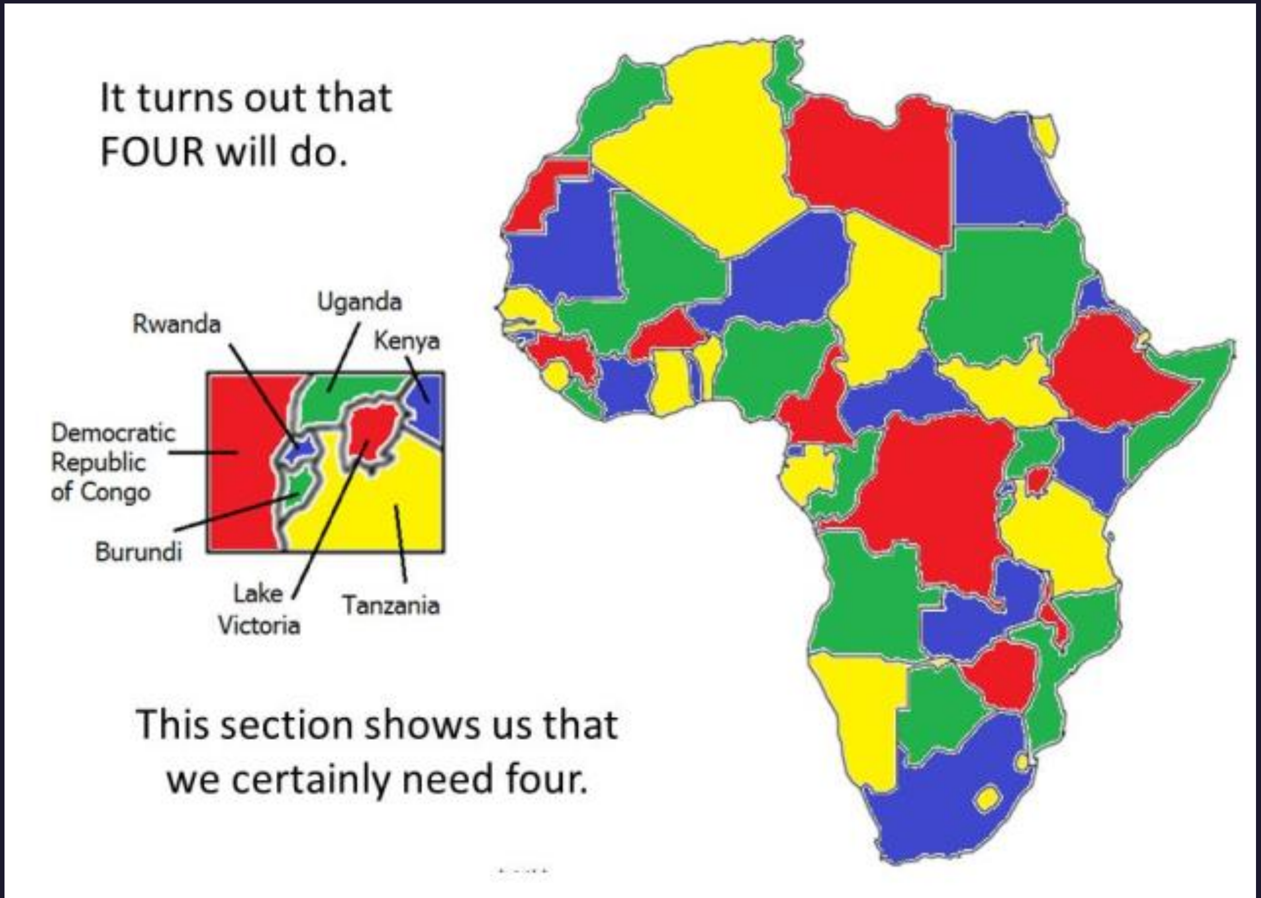
Topological Sort

- Given a DAG (Directed Acyclic Graph)
- Create a sequence such that dependencies are maintained

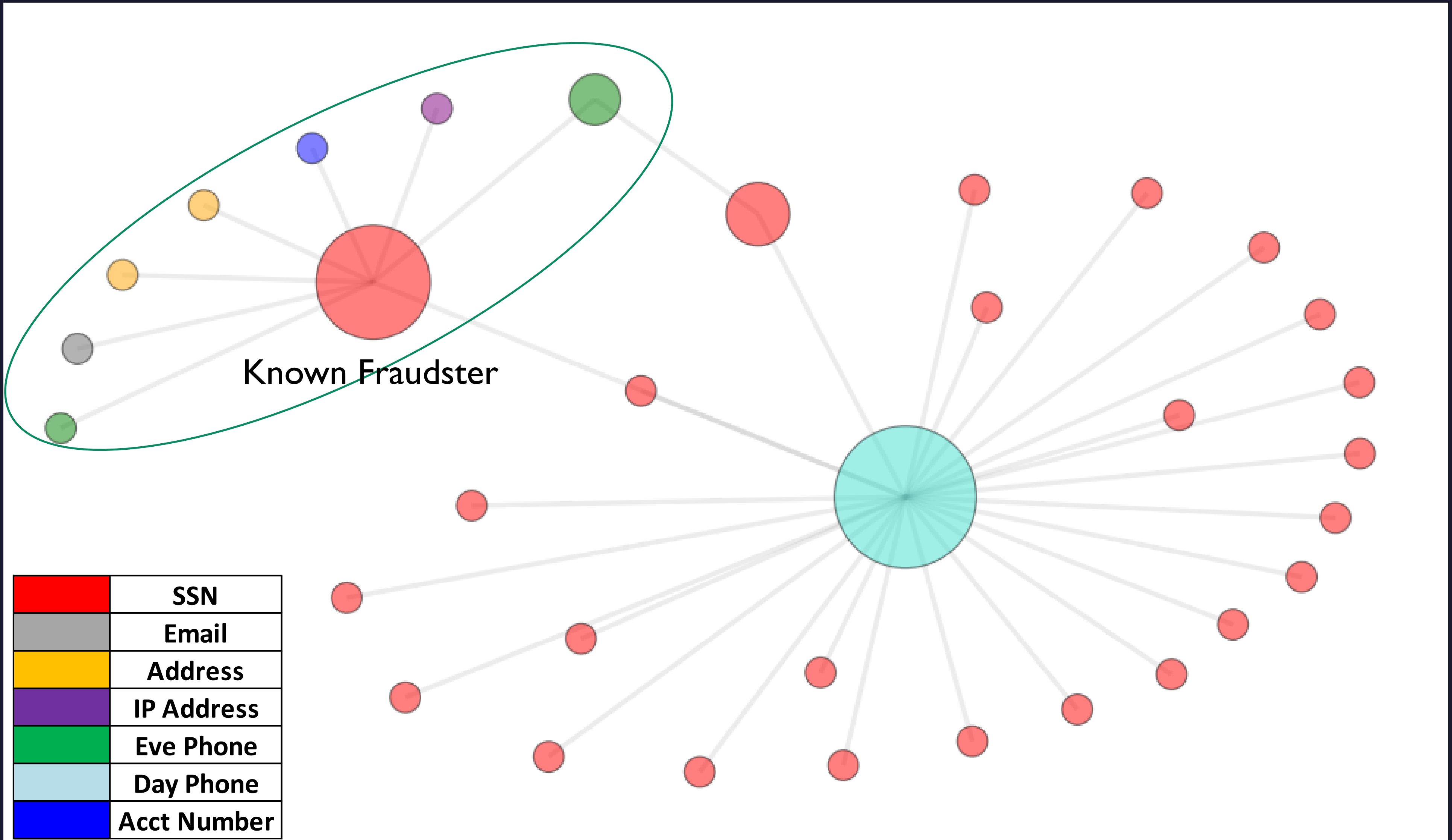


Graph Coloring

- Minimum number of colors needed so that neighboring nodes don't have the same color


It turns out that FOUR will do.
Rwanda
Uganda
Kenya
Democratic Republic of Congo
Burundi
Lake Victoria
Tanzania
This section shows us that we certainly need four.

Use Case – Identifying Fraud Rings



Questions?