Chapter 4

Graph Sampling

Problem definition

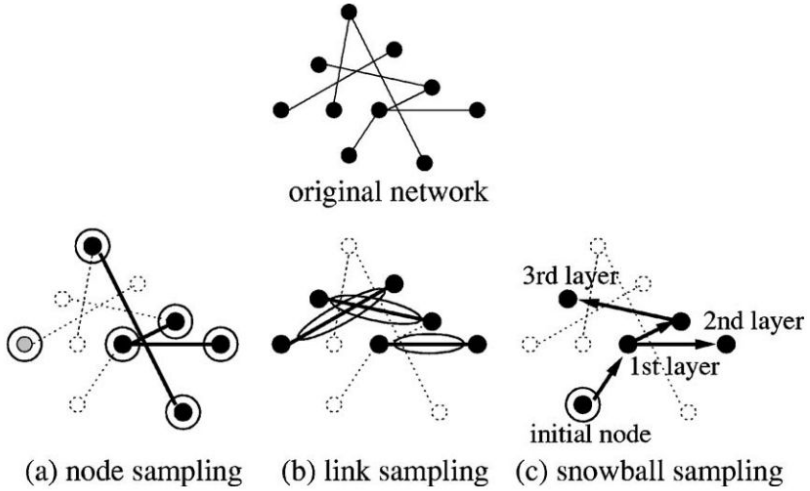
Task:
Graph sampling refers to the process of selecting a subset of nodes (or edges) from a larger graph to create a smaller, representative sample.

Why do we sample graphs?

Why and When

Some motivating reasons

- Reducing the size of graphs while aiming to retain essential properties;
- Making graph analysis feasible and efficient;
- Collect partial network views with “known” biases.



Applicative examples:

- analyzing social networks,
- understanding biological interactions,
- optimizing recommendation systems.

Large-scale graphs often pose challenges in terms of computational resources and time.

Challenges

Preserving the original graph properties,

- connectivity,
- degree distribution,
- community structure
- …

Avoiding biases:

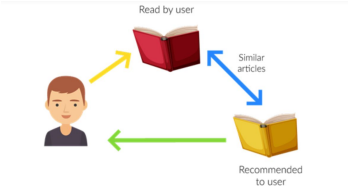
- e.g., overrepresentation of high-degree nodes

Use cases of Graph Sampling

Social Network Analysis

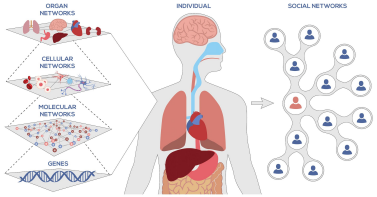
User behavior,
Information diffusion,
Influence spread

Recommendation Systems

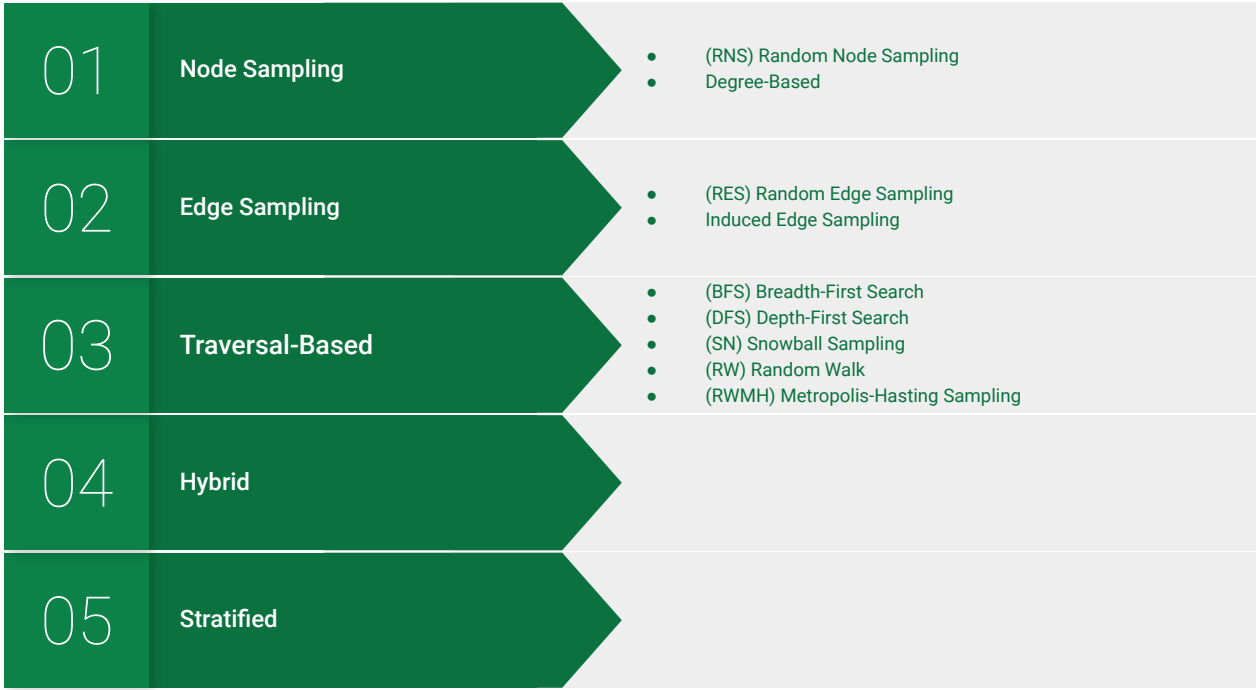


Enhance scalability/perf.

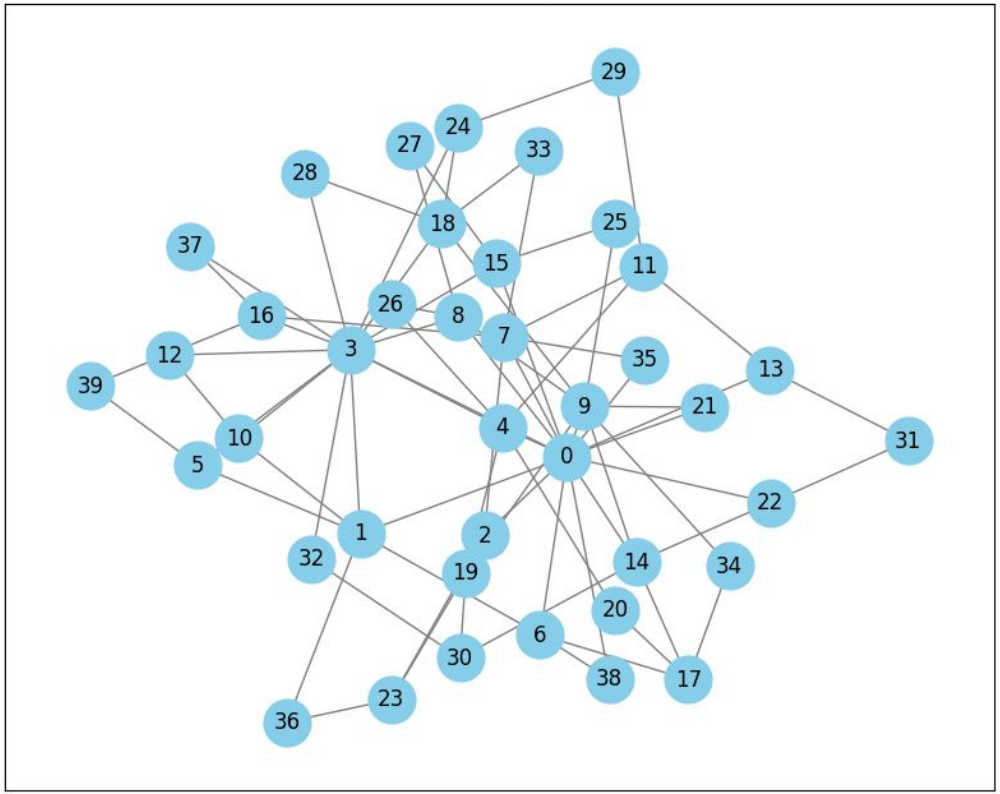
Biological Network Analysis



Protein interactions,
metabolic pathways,
gene regulatory networks



Methodologies for Graph Sampling



Running example - Barabasi-Albert (n:40, m:2)

Assumed Sampling parameters

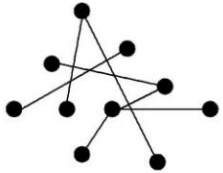
Start node: 0
Sample size: 15

Overview

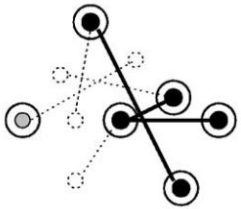
Node Sampling

Node sampling methods focus on selecting a subset of nodes from the original graph, along with the edges that connect them.

- **Random Node Sampling (RNS)**:
  Nodes are selected randomly from the graph.
- **Degree-based Node Sampling**:
  Nodes are selected based on their degree, often favoring higher-degree nodes.



original network



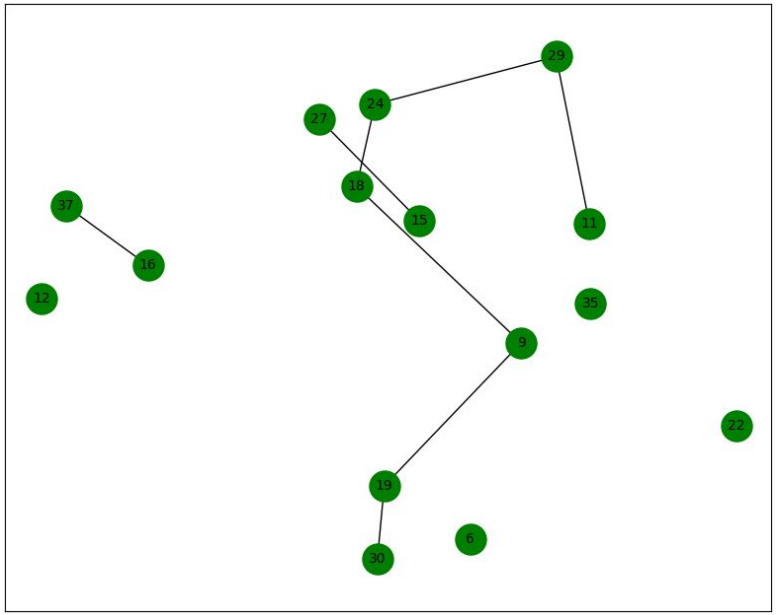
(a) node sampling

Random Node Sampling (RNS)

Randomly selecting nodes from the graph.

- Straightforward and easy to implement.
- May not always preserve important structural properties like connectivity and community structure.

RNS is often used in exploratory analysis where preserving specific properties is less critical.

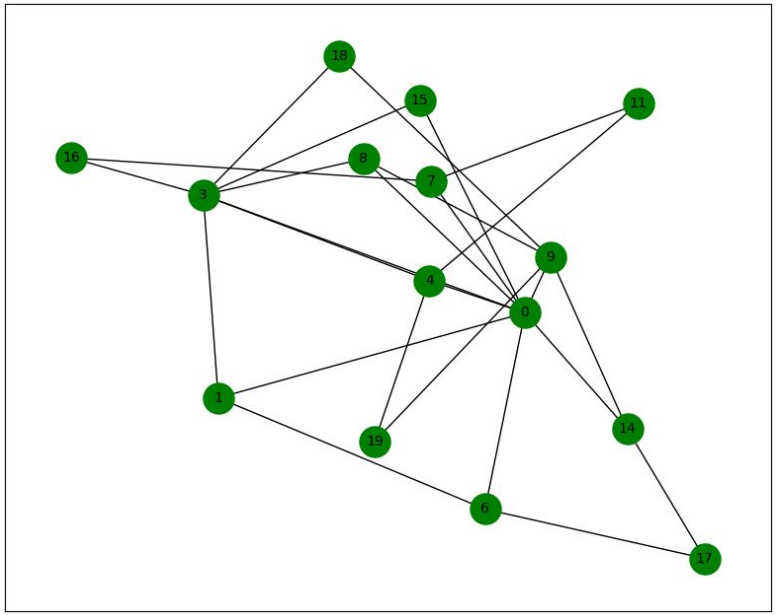


Degree-Based Node Sampling

Prioritizes nodes based on their degree, often selecting nodes with higher degrees more frequently.

- Preserve the graph's degree distribution;
- can introduce biases by overrepresenting central nodes.

It is useful when the focus is on analyzing hub structures and central nodes in the graph.

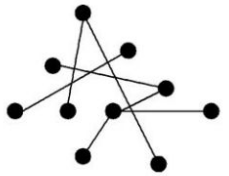


Overview

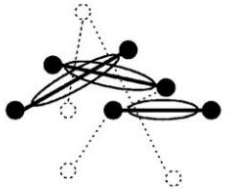
Edge Sampling

Edge sampling methods focus on selecting a subset of edges from the original graph, possibly including the nodes connected by these edges.

- **Random Edge Sampling (RES)**:
  Edges are selected randomly from the graph.
- **Induced Edge Sampling**:
  Edges are selected such that the resulting subgraph includes all nodes connected by the sampled edges.



original network



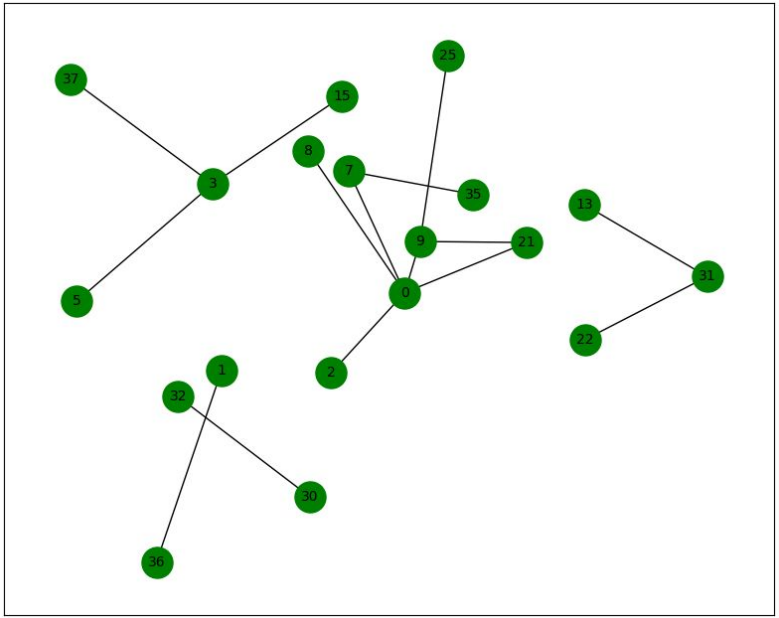
(b) link sampling

Random Edge Sampling (RES)

Randomly selecting edges from the graph.

- Lead to the inclusion of nodes with various degrees, providing a diverse sample.
- It may disrupt the graph's connectivity and community structure.

RES is useful when the goal is to analyze general edge-related properties without focusing on specific node characteristics.

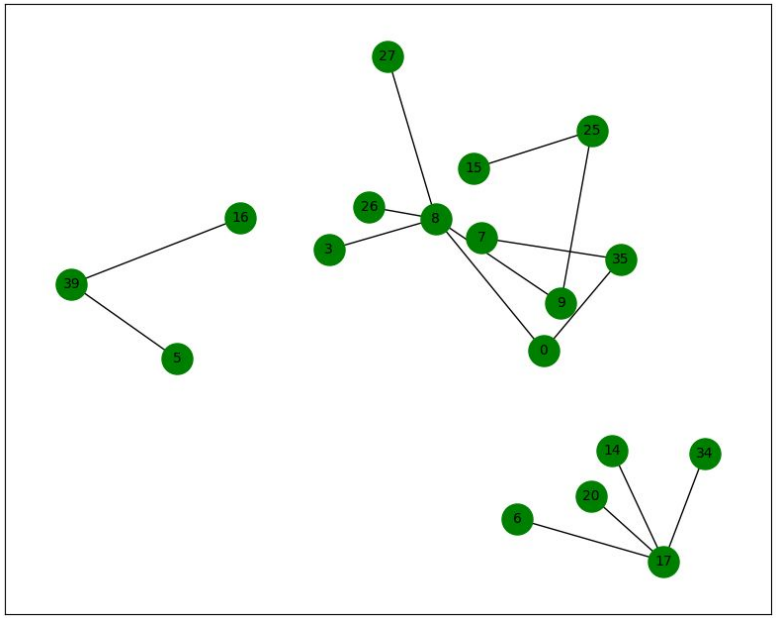


Induced Edge Sampling

Selects edges such that all nodes connected by the sampled edges are included in the subgraph.

- Preserve local structures and community formations.

It is particularly useful when studying the connectivity and clustering properties of the graph.

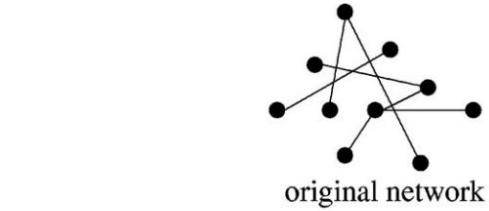


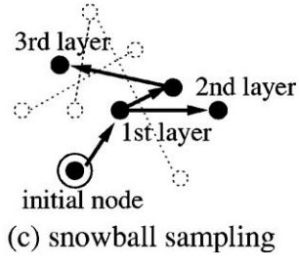
Overview

Traversal-Based Sampling

Traversal-based sampling methods navigate through the graph to select nodes and edges

- **Breadth-First Sampling (BFS)**:
  Nodes are explored level by level.
- **Depth-First Sampling (DFS)**:
  Nodes are explored by diving deep into each branch before backtracking.
- **Snowball Sampling**:
  Starts from an initial set of nodes and expands outward, mimicking a snowball effect.



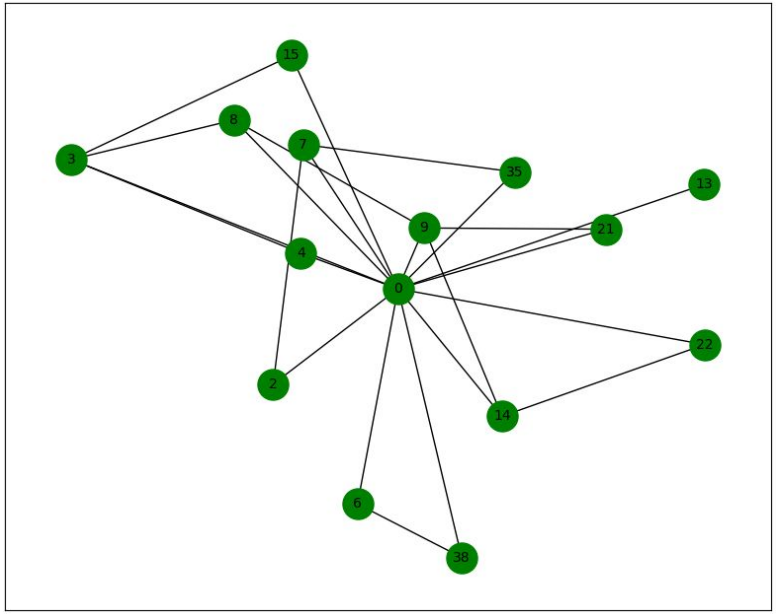


Breadth-First Sampling (BFS)

Explores the graph level by level, starting from an initial node and expanding outwards.

- Effective for sampling connected components and preserving local structures.
- May introduce biases towards densely connected regions.

BFS is useful for tasks that require understanding the immediate neighborhood around nodes.

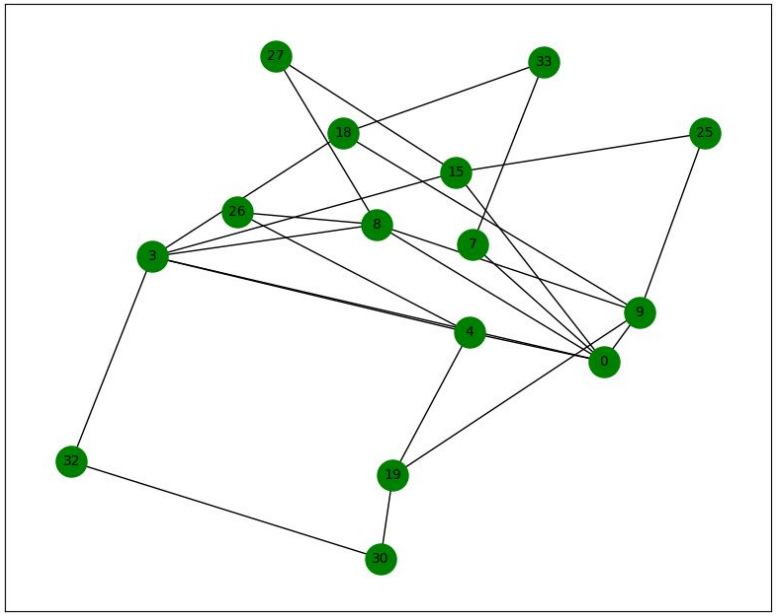


Depth-First Sampling (DFS)

Explores the graph by diving deep into each branch before backtracking.

- Capture long paths and deeper structures within the graph.
- May miss out on broader, shallower structures.

DFS is suitable for tasks that require analyzing deep hierarchical or chain-like structures.

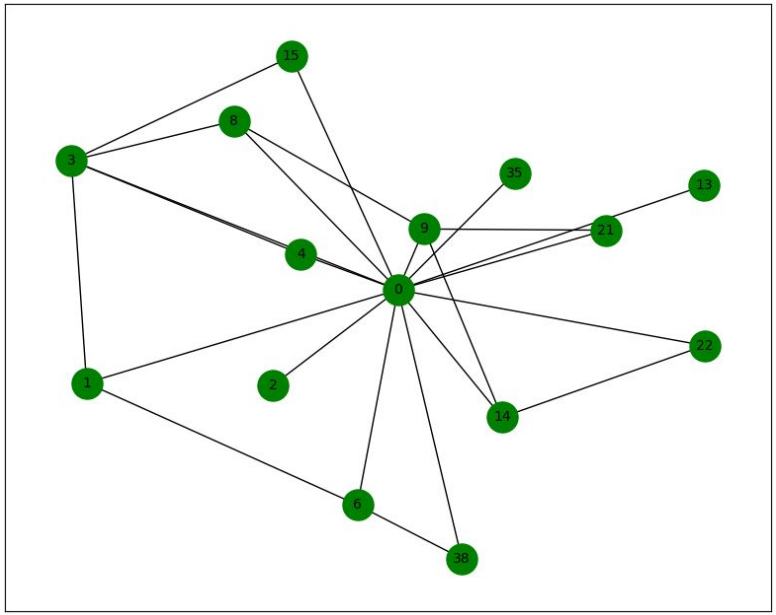


Snowball Sampling

Starts from a set of initial nodes and expands outward by exploring their neighbors, mimicking a snowball effect.

- Preserves local neighborhood relationships;
- Can introduce biases towards initially chosen nodes.

Useful for studying social networks and community structures.

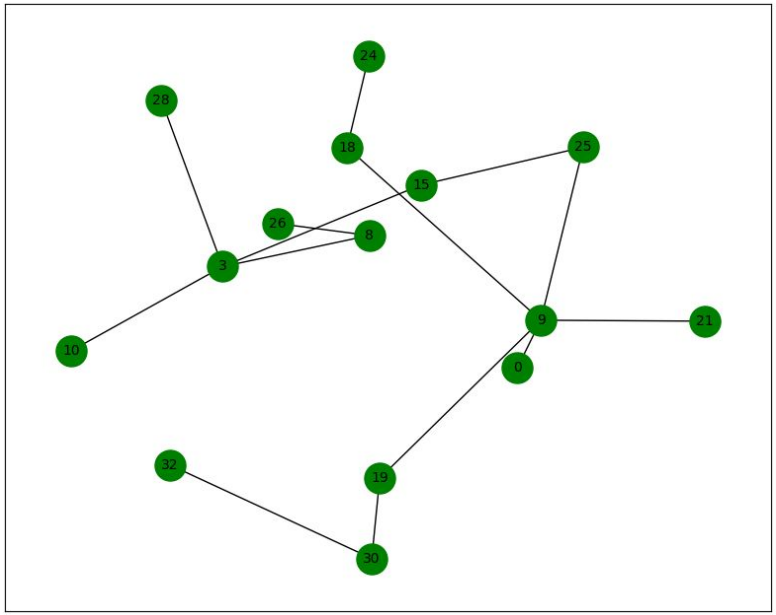


Random Walk Sampling

Involves selecting nodes by performing a random walk on the graph.

- Nodes are visited randomly based on their connections.
- Capture the graph's structural properties.

It is often used in network analysis and sampling from large, complex graphs.

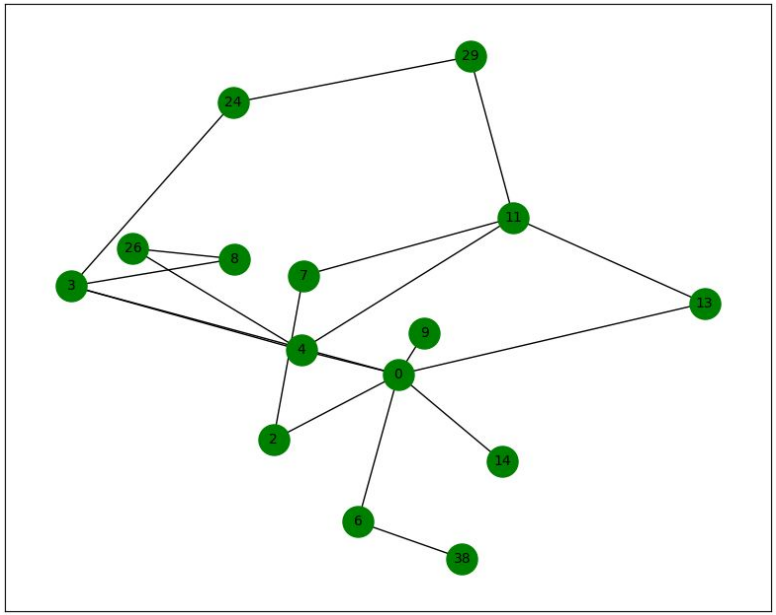


Metropolis-Hastings Sampling

Extends the random walk approach by using a probabilistic method to accept or reject the next node, ensuring a more uniform sample.

- Helps mitigating biases in random walks

It is useful for tasks requiring a representative sample of the entire graph.



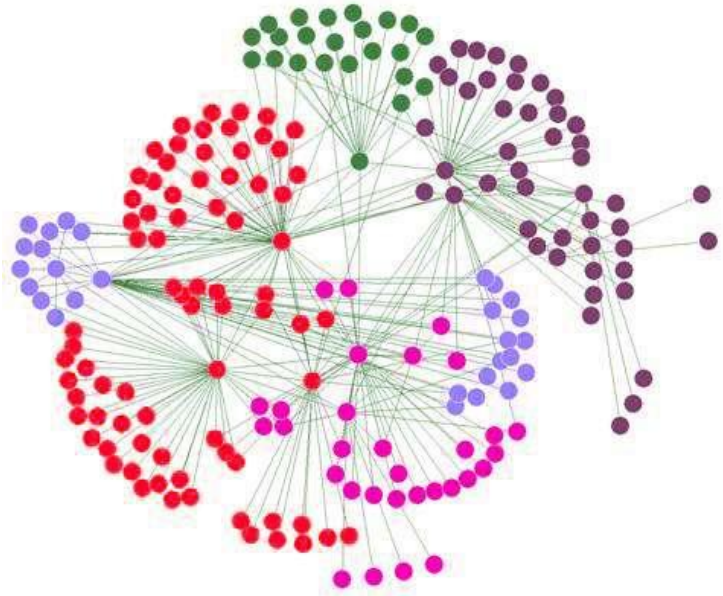
Overview

Hybrid Sampling

Hybrid sampling methods combine different techniques to leverage the strengths of each approach.

These methods aim to achieve a balanced sample that preserves various graph properties.

- e.g., combining random node sampling with induced edge sampling can help maintain connectivity and degree distribution.



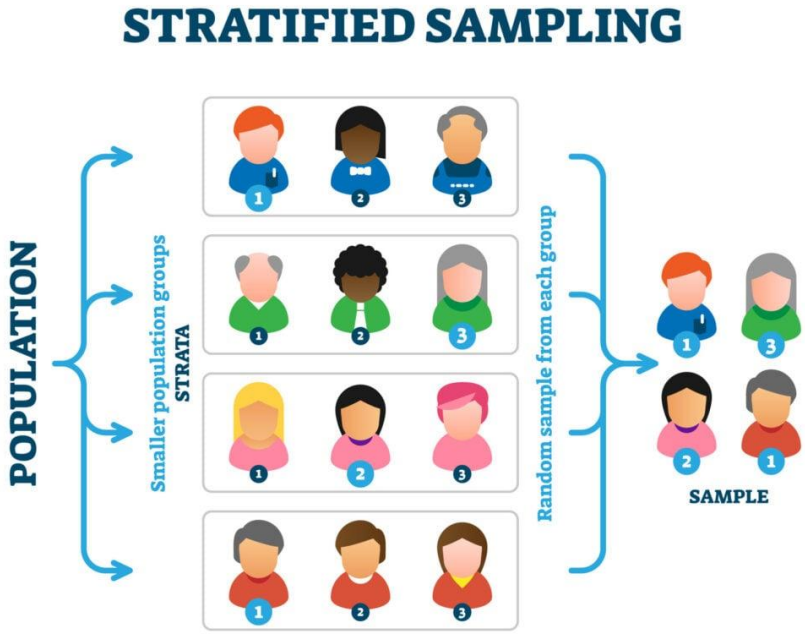
Overview

Stratified Sampling

Stratified Sampling involves dividing the graph into strata based on node or edge attributes and then sampling from each stratum.

It ensures that different segments of the graph are proportionally represented in the sample.

- e.g., It is useful for preserving diversity and ensuring that specific subgroups are not overlooked.



Impact on Sampled Topology

Impact on Structural Properties

Sampling methods impact various structural properties of the graph. E.g.,

- **Degree Distribution**:
  The frequency of nodes with different degrees.
- **Clustering Coefficient**:
  The tendency of nodes to form tightly knit groups.
- **Path Length and Diameter**:
  The average and maximum distances between nodes.

Common Sampling Biases

Degree Bias:
Overrepresentation of high-degree nodes.

Structural Bias:
Distortion of the graph's structural properties.

Mitigating these biases involves careful selection of sampling techniques and combining multiple methods to balance their effects.

Understanding these biases is crucial for interpreting the results accurately.

Impact on Structural Properties

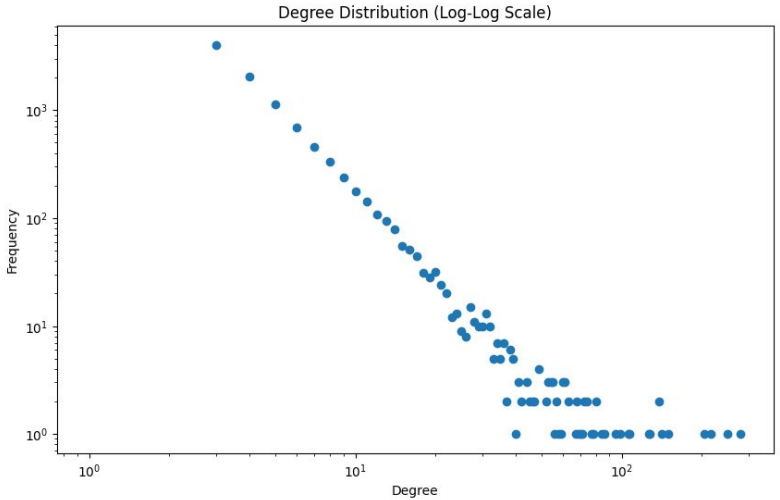
Degree Distribution

Degree distribution reflects the number of nodes having each possible degree.

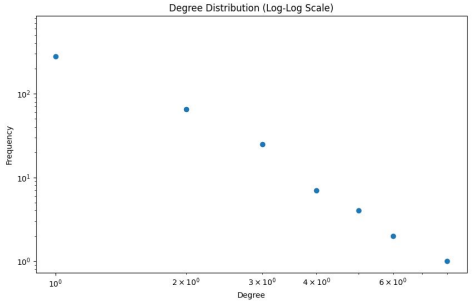
Sampling can alter this distribution, especially if certain nodes are more likely to be included.

- E.g., degree-based sampling may overrepresent high-degree nodes, skewing the distribution.

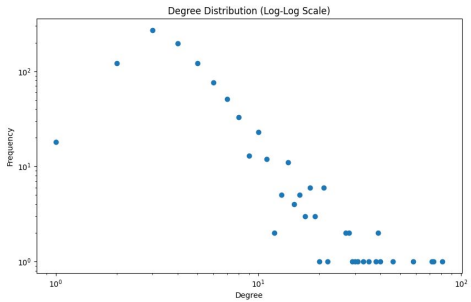
Ensuring that the sample's degree distribution closely matches the original graph's distribution is crucial for accurate analysis.



G = nx.barabasi_albert_graph(10000, 3)

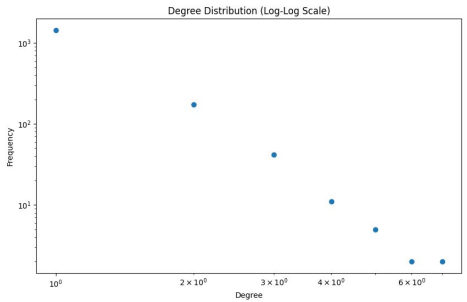


Random Node Sampling

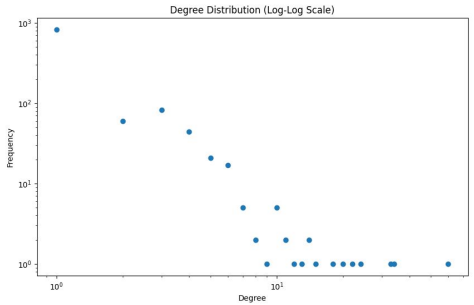


Degree Node Sampling

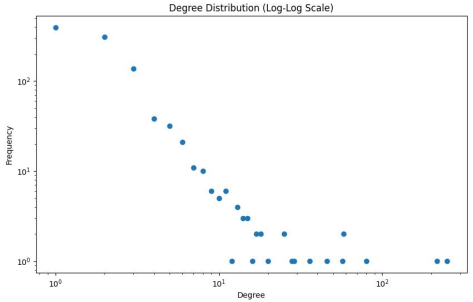
Sampling parameters
Starting node: 0
Sample Size: 1000



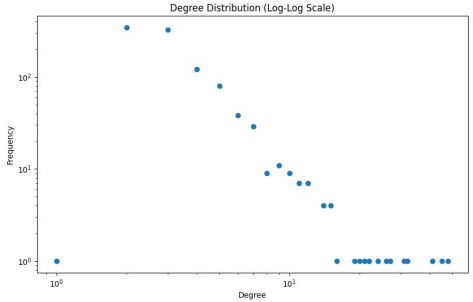
Random Edge Sampling



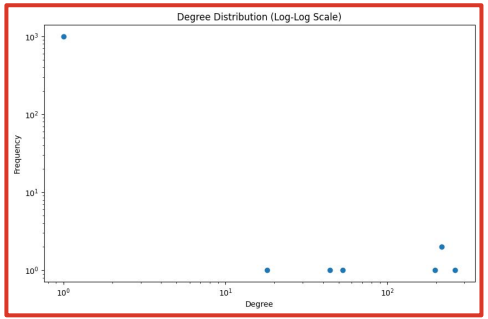
Induced Edge Sampling



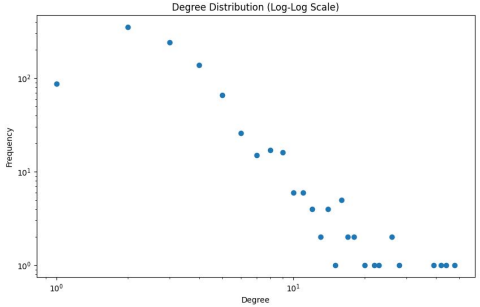
BFS



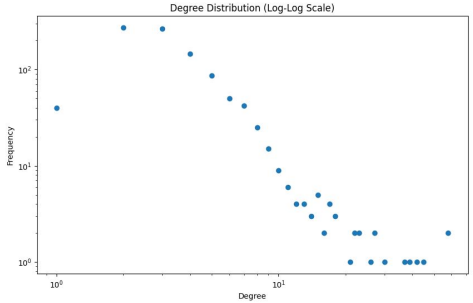
DFS



Snowball



Random Walk



Metropolis-Hasting

Impact on Structural Properties

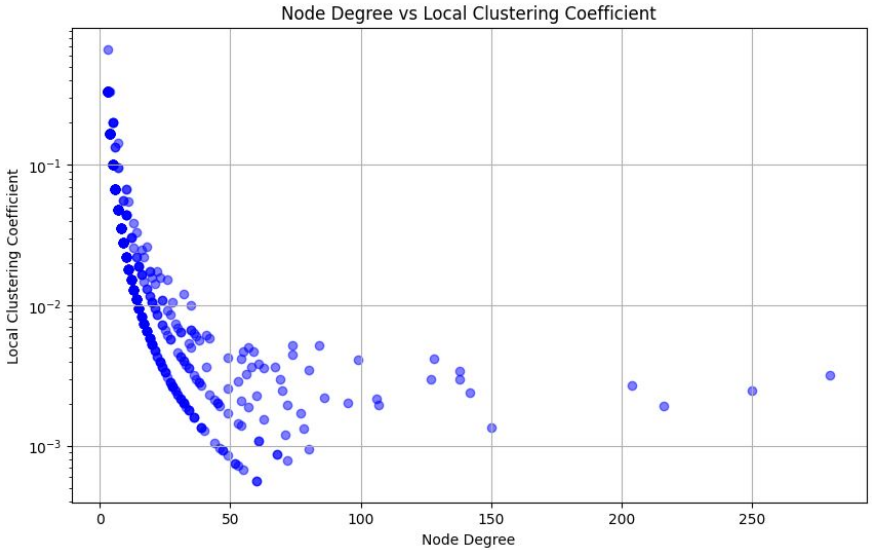
Clustering Coefficient

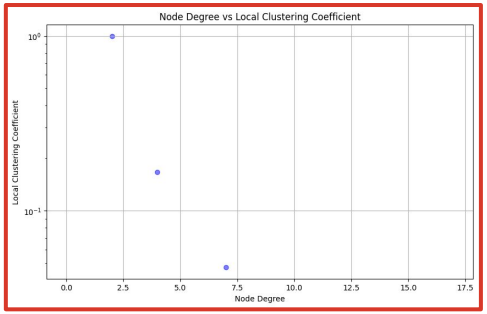
The clustering coefficient measures the likelihood that a node's neighbors are also connected.

Sampling methods can affect this coefficient by disrupting local structures.

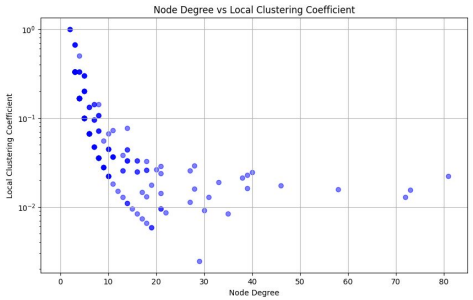
- E.g., random edge sampling may reduce the clustering coefficient if key edges forming triangles are not included.

Analyzing how different sampling methods impact clustering helps in selecting appropriate techniques for preserving local connectivity.



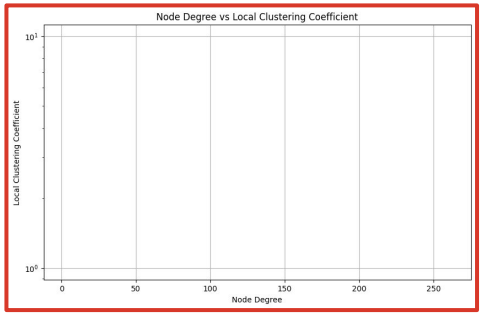


Random Node Sampling

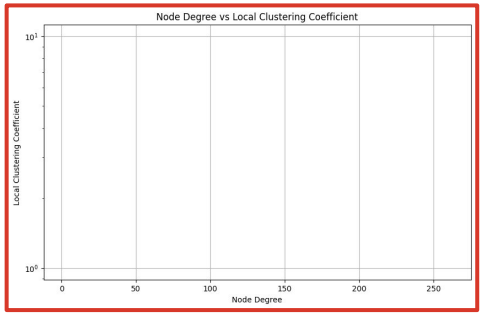


Degree Node Sampling

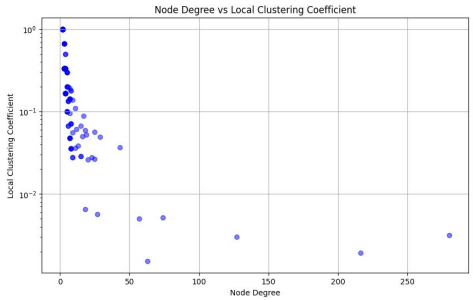
Sampling parameters
Starting node: 0
Sample Size: 1000



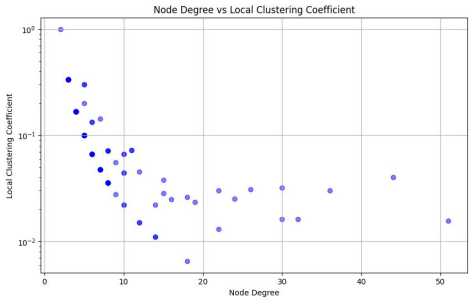
Random Edge Sampling



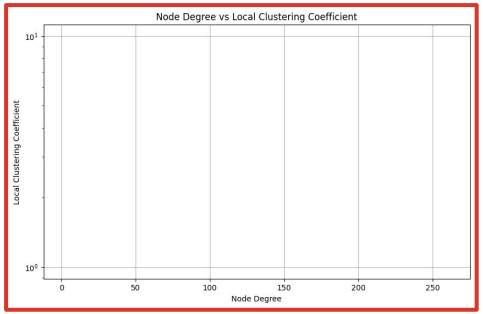
Induced Edge Sampling



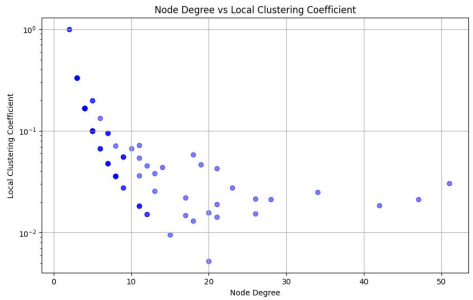
BFS



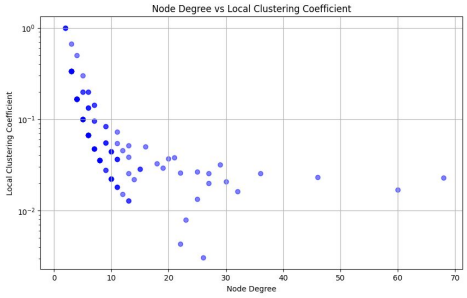
DFS



Snowball



Random Walk



Metropolis-Hasting

Impact on Structural Properties

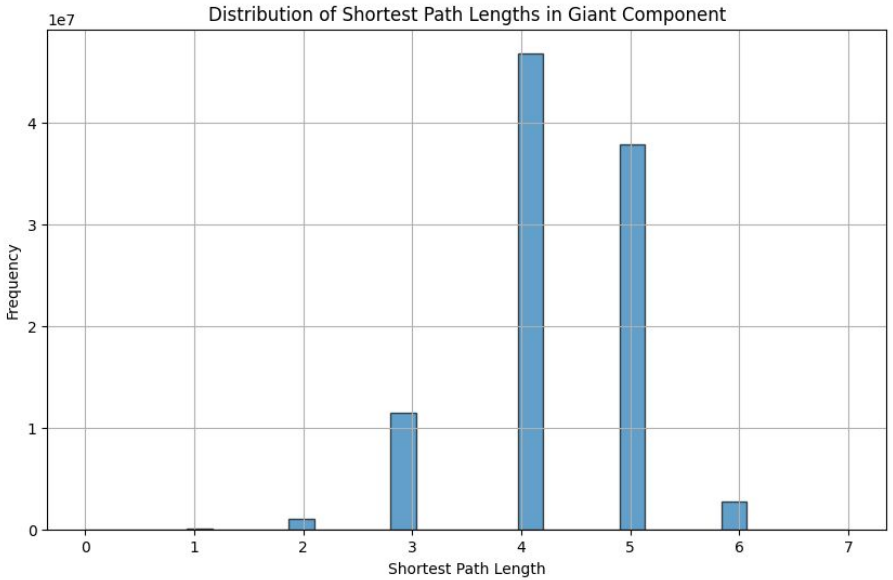
Distances

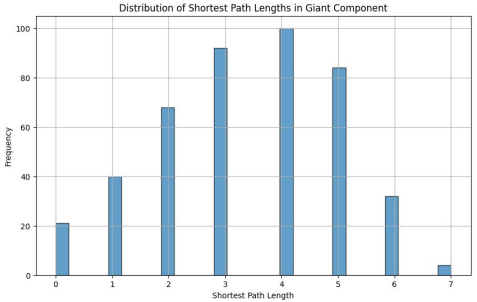
Path length refers to the average distance between pairs of nodes.

Sampling can shorten or lengthen these distances depending on the method used.

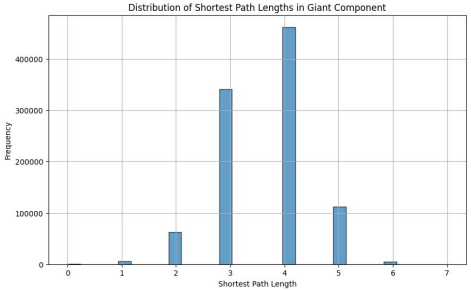
- BFS might preserve short path lengths within local clusters;
- RNS might disrupt longer paths.

Understanding these impacts is crucial for applications involving shortest paths and network navigation.



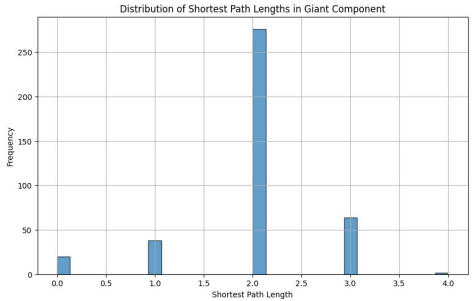


Random Node Sampling

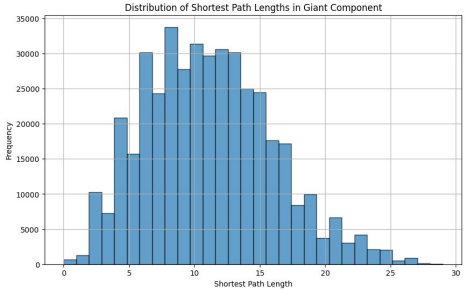


Degree Node Sampling

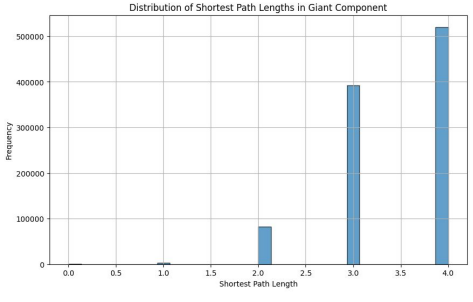
Sampling parameters
Starting node: 0
Sample Size: 1000



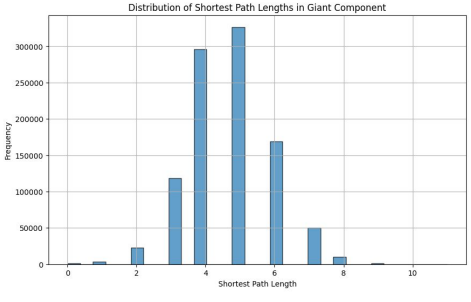
Random Edge Sampling



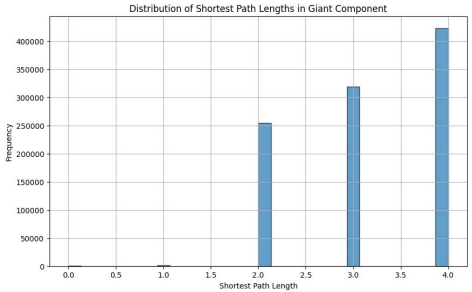
Induced Edge Sampling



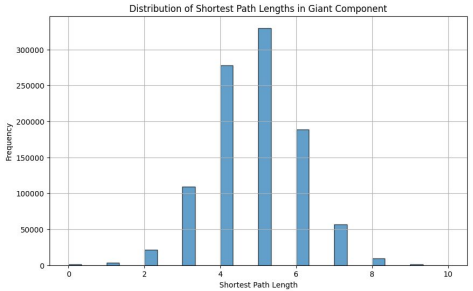
BFS



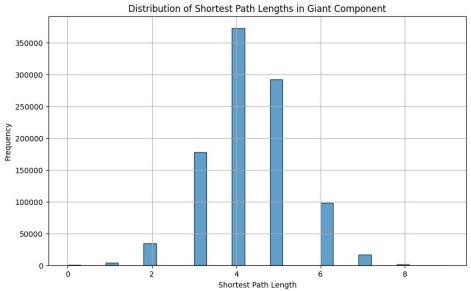
DFS



Snowball



Random Walk



Metropolis-Hasting

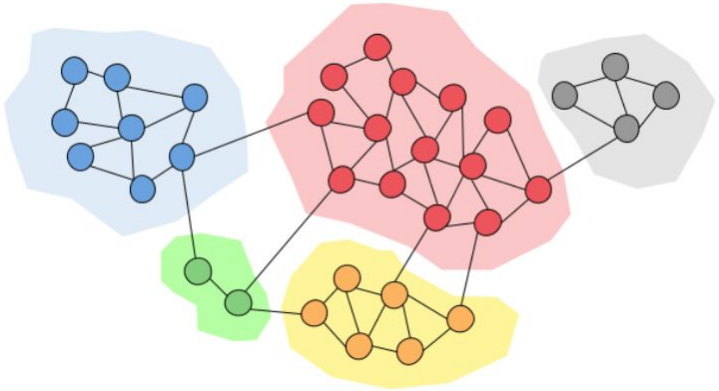
Impact on Meso-scale Properties

Community Structure

Communities are groups of nodes that are more densely connected internally than with the rest of the graph.

Sampling can impact community detection by including or excluding critical nodes and edges.

Ensuring that sampling methods retain community structures is essential for applications focusing on social network analysis and clustering.



What if I have only partial graph knowledge?

Issue: Data Collection

When dealing with partially known data certain graph sampling algorithms are more suitable than others

- **Snowball Sampling**
  It leverages the partial data by expanding outward from initial known nodes, making it ideal when only parts of the network are accessible.

- **Random walk/Metropolis-Hasting**
  Explore the network based on available data at each step, making it effective when only parts of the network are accessible at any time.

Example of application scenarios

- OSN data collection via APIs
- Epidemiology studies relying on confirmed contacts
- Protein-Protein interaction from incomplete observational data
- Urban mobility from digital traces

Does Sampling affect HP?

Sampling & Hypothesis

Representation of the Original Graph

Bias: Different sampling methods can introduce different types of bias.

Coverage: Incomplete sampling may miss important nodes or edges, affecting the representativeness of the sample

Statistical Power

Sample Size: The number of nodes and edges in the sampled graph can affect the statistical power of hypothesis tests.

- Smaller samples may lead to less reliable estimates and higher variability.

Sampling Error: The variability introduced by the sampling can affect the accuracy and precision of hypothesis tests.

Inference Methods

Estimator Bias: Sampling can affect the bias and variance of estimators used in hypothesis testing.

- E.g., degree distribution estimators may be biased if the sampling method is not uniform

Corrective Measures: Techniques (e.g., reweighting) can correct for the sampling biases. However, they also come with their own assumptions and limitations.

Examples of impacts on Hypotheses

Social Networks

Sampling might affect hypotheses related to influence spread.

If influential nodes are undersampled, the estimated spread of influence might be underestimated.

Epidemiology

In disease transmission networks, sampling can affect hypotheses about the rate of spread or the identification of super-spreaders.

Missing key nodes or connections could lead to incorrect conclusions about disease dynamics.

Infrastructure Networks

In power grids or transportation networks, sampling might affect hypotheses about network resilience.

If critical nodes or edges are missed, the network might appear more resilient than it actually is.

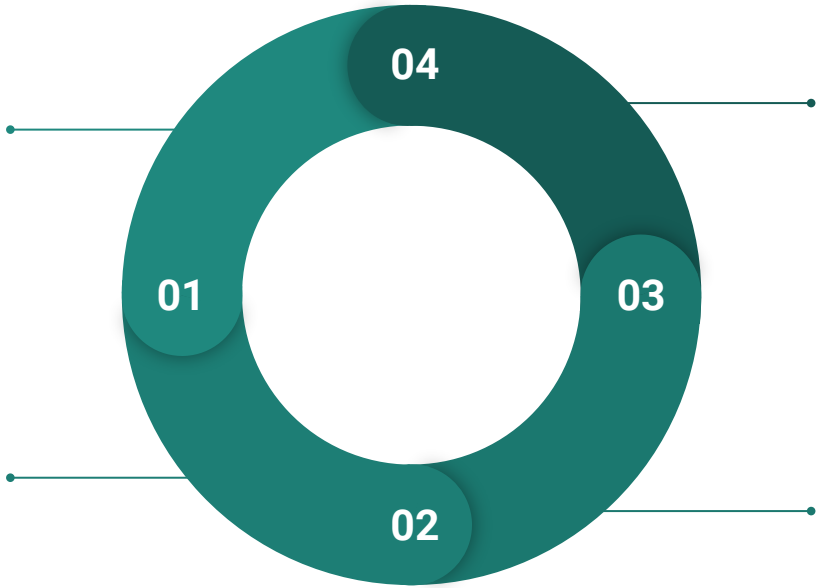
Mitigating the Effects of Sampling

Validation

Use multiple sampling methods and compare results to assess the robustness of hypotheses.

Adjustment Techniques

Apply statistical adjustments to account for known biases in the sampling method.



Simulation Studies

Conduct simulation studies to understand the potential impacts of different sampling methods on hypothesis testing.

Larger Samples

Whenever possible, increase the sample size to reduce sampling error and improve representativeness.

Chapter 4

Conclusion

Take Away Messages

1. Sampling allows to simplify a network preserving some characteristics
2. Each strategy provides different guarantees

Suggested Readings

- The Atlas for the aspiring Network Scientist (Ch 25)
- https://user.informatik.uni-goettingen.de/~ychen/papers/graph_sampling_simplex11.pdf
- https://arxiv.org/pdf/1308.5865

What's Next

Chapter 5: Who needs data?