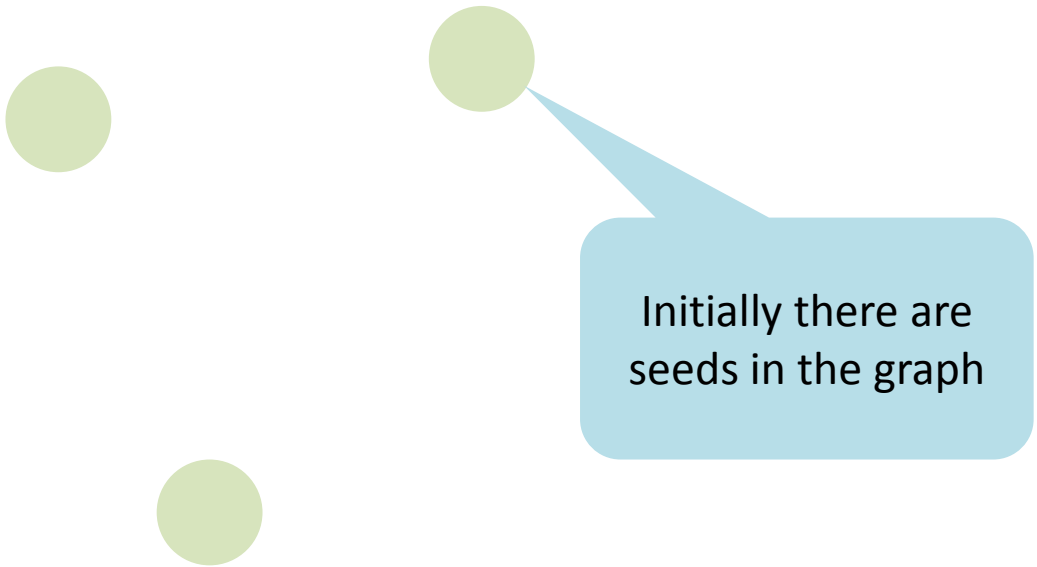
HW3: Social Network Sampling

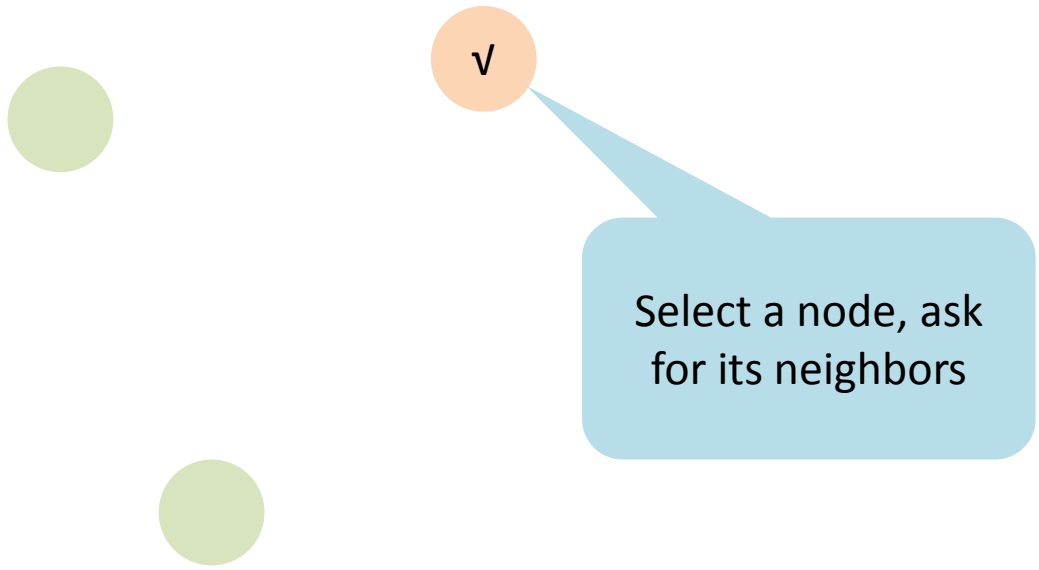
Scenario

- Given a social network as a graph, sample subgraphs to estimate the properties (degree distribution, centrality, etc.) of the original graph
- Note that you don’t have the full view of the network
  - Imagine that Facebook or Twitter that does not release their whole social network
  - We develop a crawler to browse the network starting from our own accounts

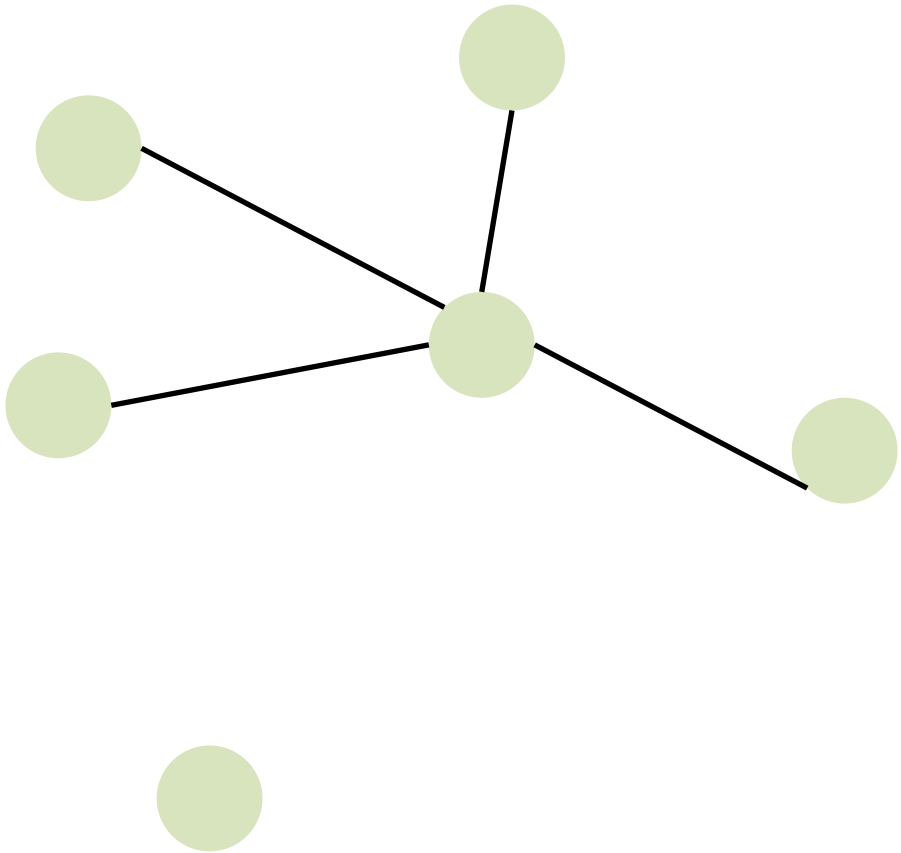
Example



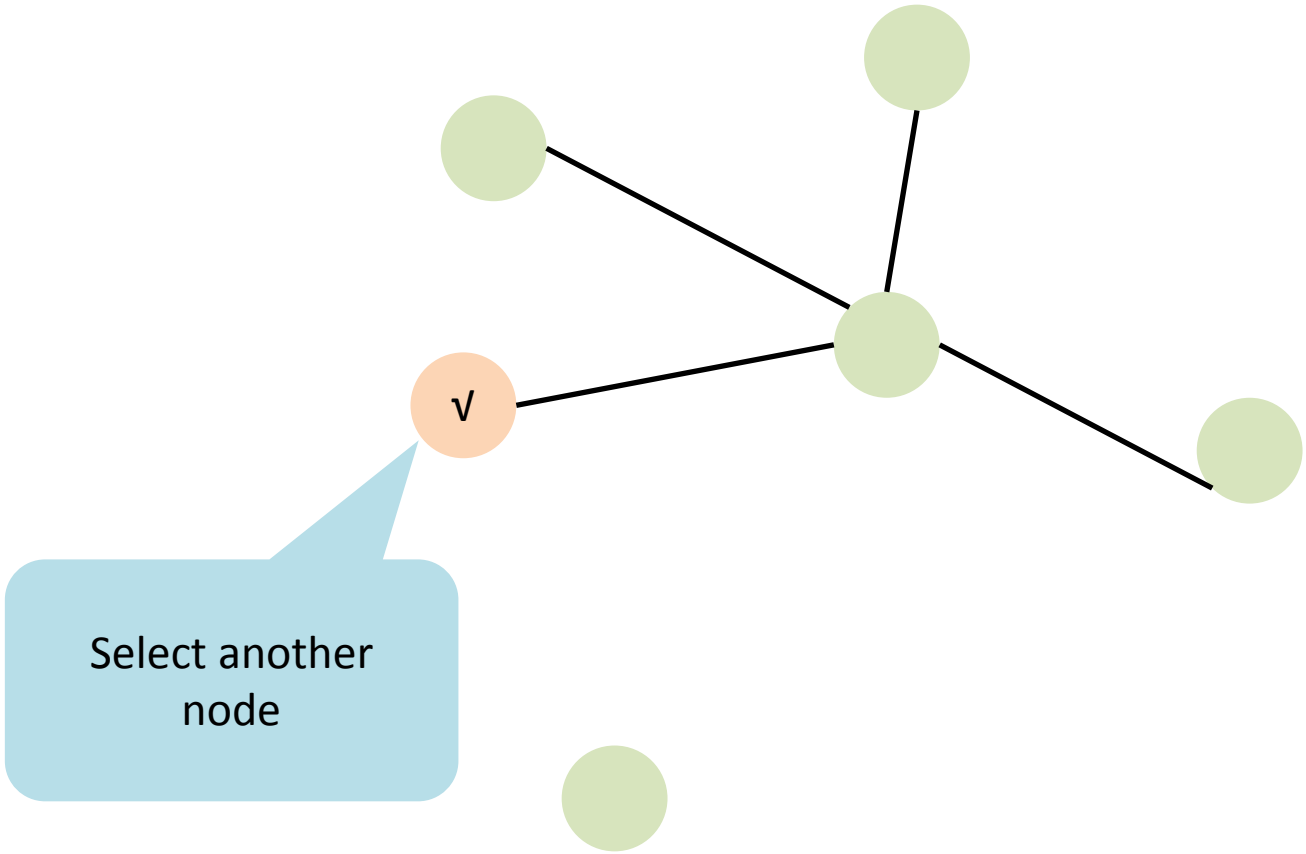
Example



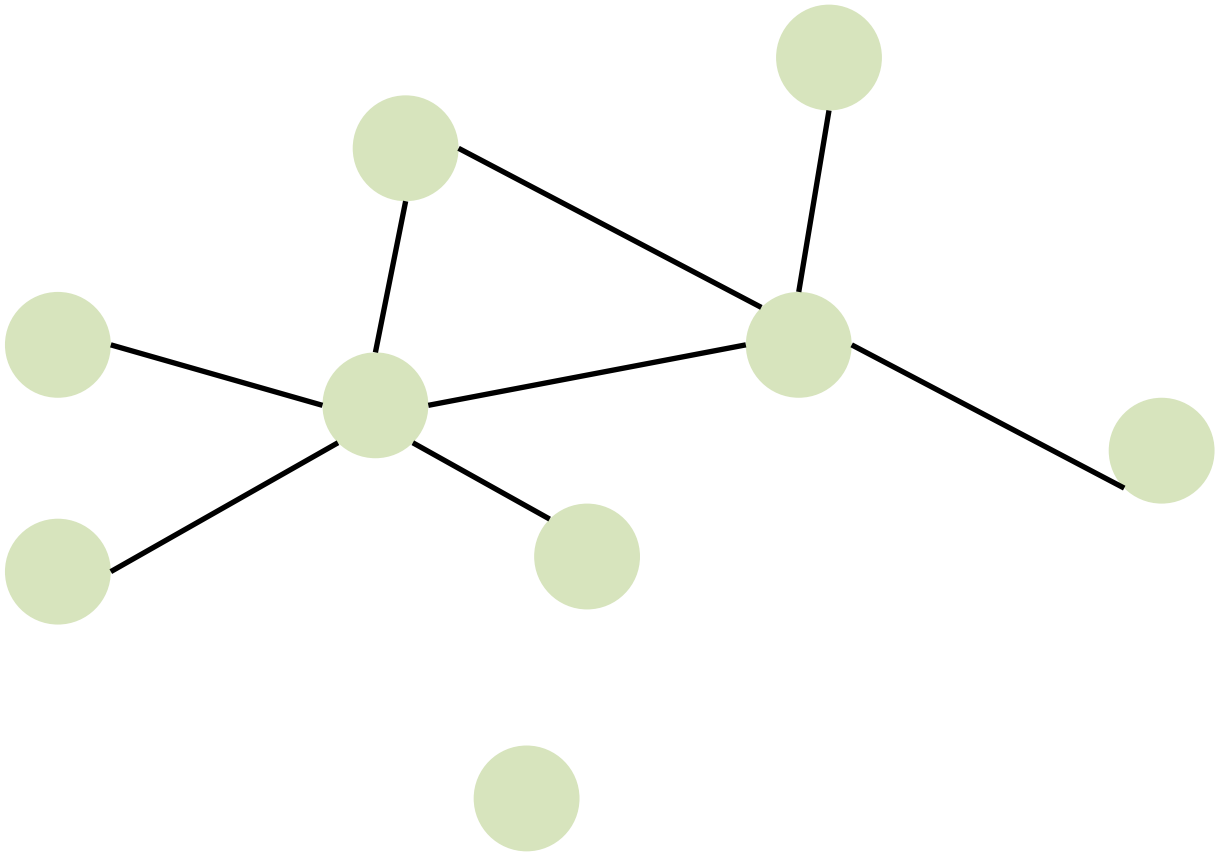
Example



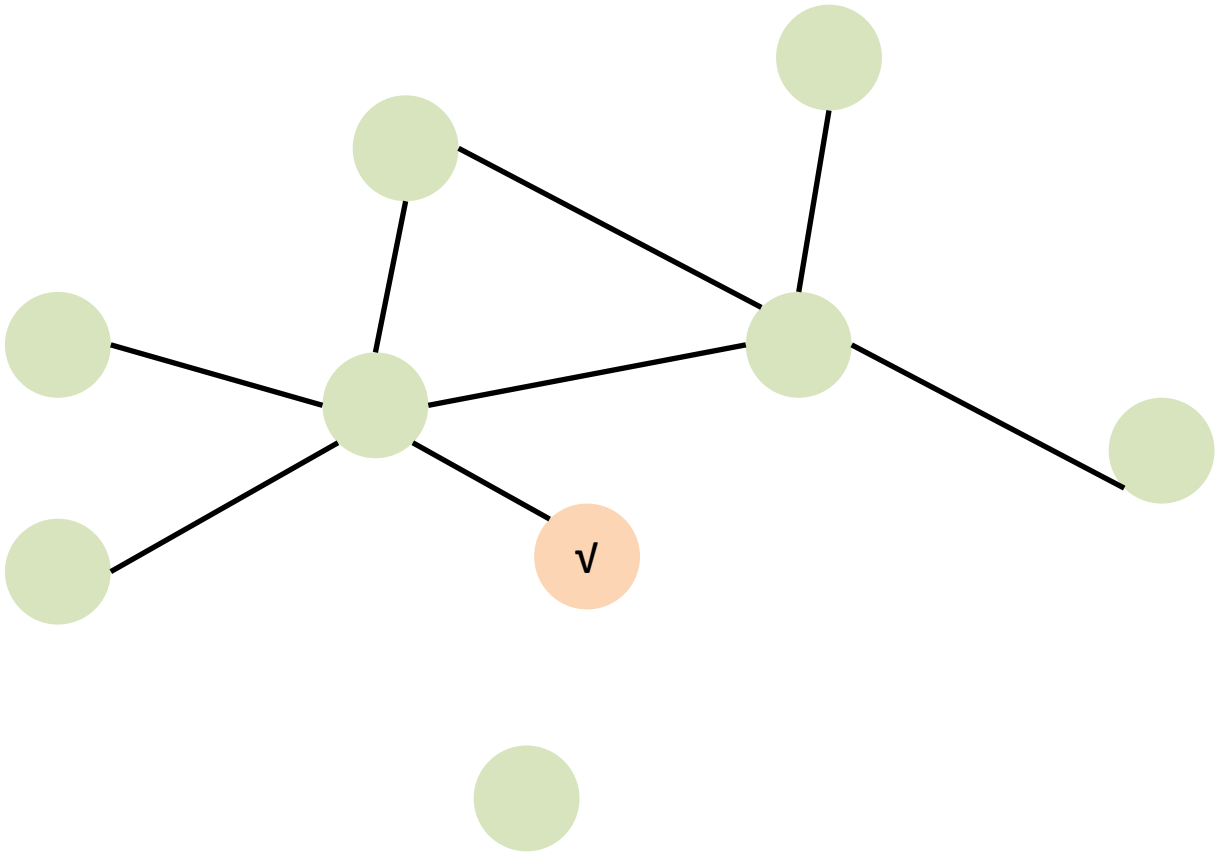
Example



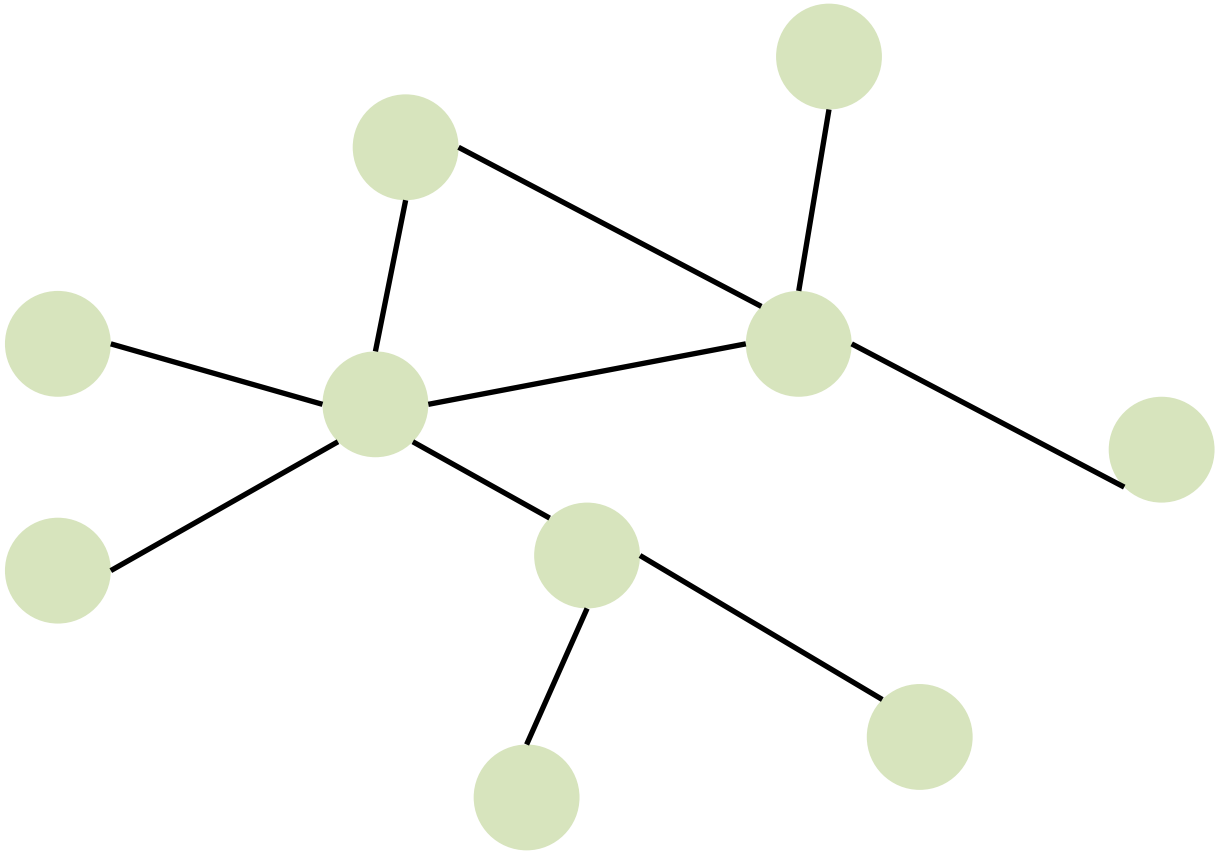
Example



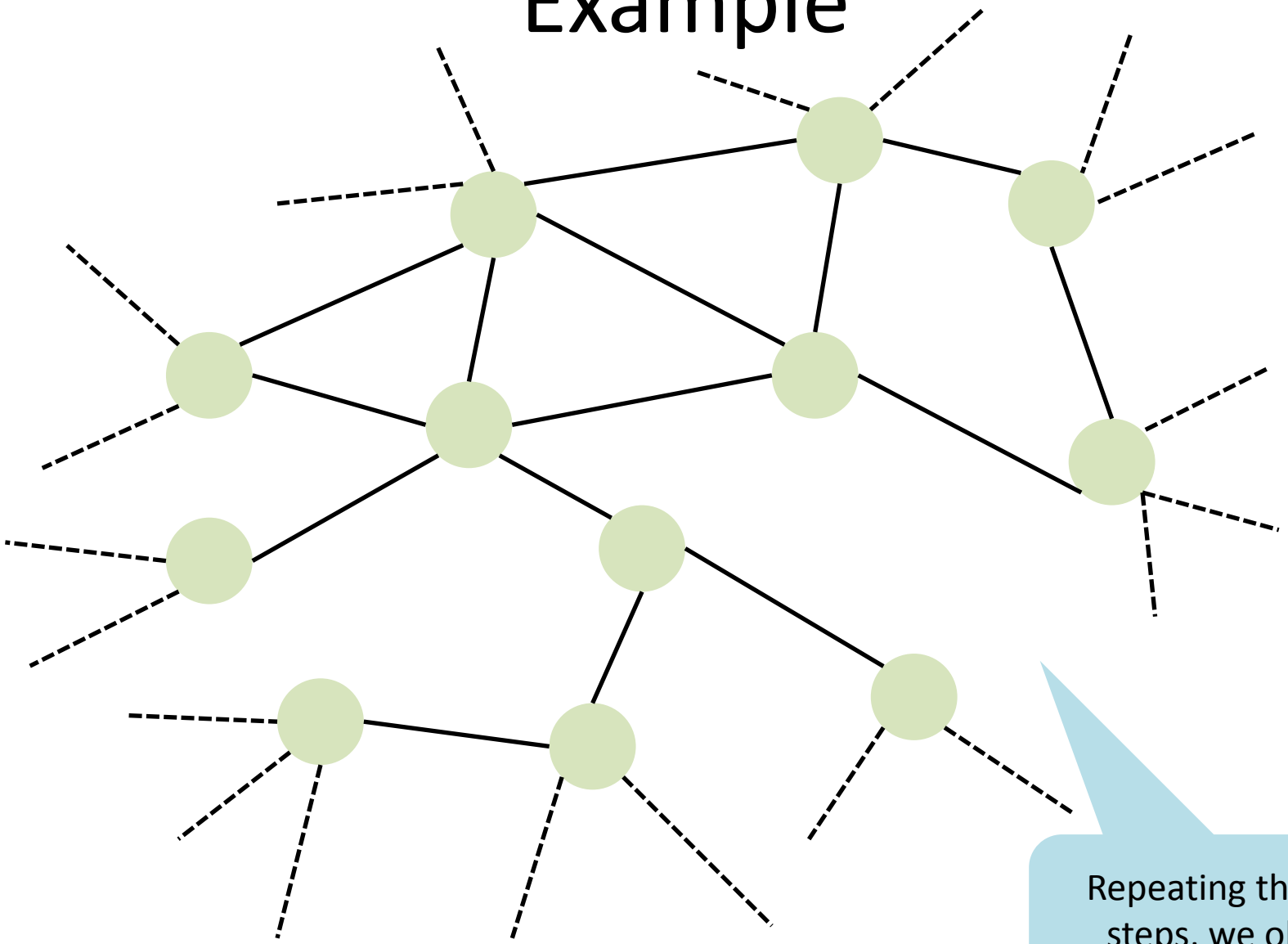
Example



Example



Example



Graphs for HW3

- The given graphs is
  - undirected
  - connected
- It is a heterogeneous network
  - There are k_V attributes for nodes
    - $1 \leq k_V \leq 5$
    - Each attribute is represented by categorical or numerical integers
  - There are k_E attributes for edges
    - $0 \leq k_E \leq 1$
    - Each attribute is represented by categorical or numerical integers

Query System

- HTTP GET request
  - http://140.112.31.186/SNA2014/hw3/query.php?team=XXXX&node=YYYY
  - XXXX: Password of your team
    - We have different node id assignment for each team
  - YYYY (optional): Node id whose neighbors you would like to know
    - If not assigned, you will be provided the seed subgraph

Query System Usage

- Example
  - http://140.112.31.186/SNA2014/hw3/query.php?team=XXXX&node=2014
  - Let $k_V = 2, k_E = 1$

```
26                  # Team id
10                  # The ith query (query count +1)
2014 11 2 5         # Queried node id , degree d, k_V node attributes
67 45 3 78 6        # Neighbor id, degree, k_V node attributes ,
                    # k_E edge attributes between node 2014 and 67
28 5 1 32 2
...
999 1 0 45 9
```

d lines represents d neighbors

Seed Subgraph

- Example
  - http://140.112.31.186/SNA2014/hw3/query.php?team=XXXX

```
26                  # Team id
10                  # The ith query (count not changed)
2 1                 # k_V, k_E
100                 # Number n_s of seed nodes
12 34 7 2           # Node id , degree d, k_V node attributes
...
99 54 2 4
12 13 0             # Edge (12, 13), k_E edge attribute
...
97 99 1
```

n_s nodes

The following are edges

- The seed subgraph will be updated at 0:00 everyday
- The query count will be reset 0 at 0:00 everyday

Special Responses

- If you can no longer query nodes
  - For private graph only; allowed number of queries is limited
  - Only one line “-1”

```
-1
```

- Nodes with degree 0
  - The node you queried does not exist (our graphs are connected; every node has at least degree 1)

```
26              # Team id
1               # The ith query (query count +1)
244 0           # Degree 0; node 244 does not exist
```

Query System Schedule

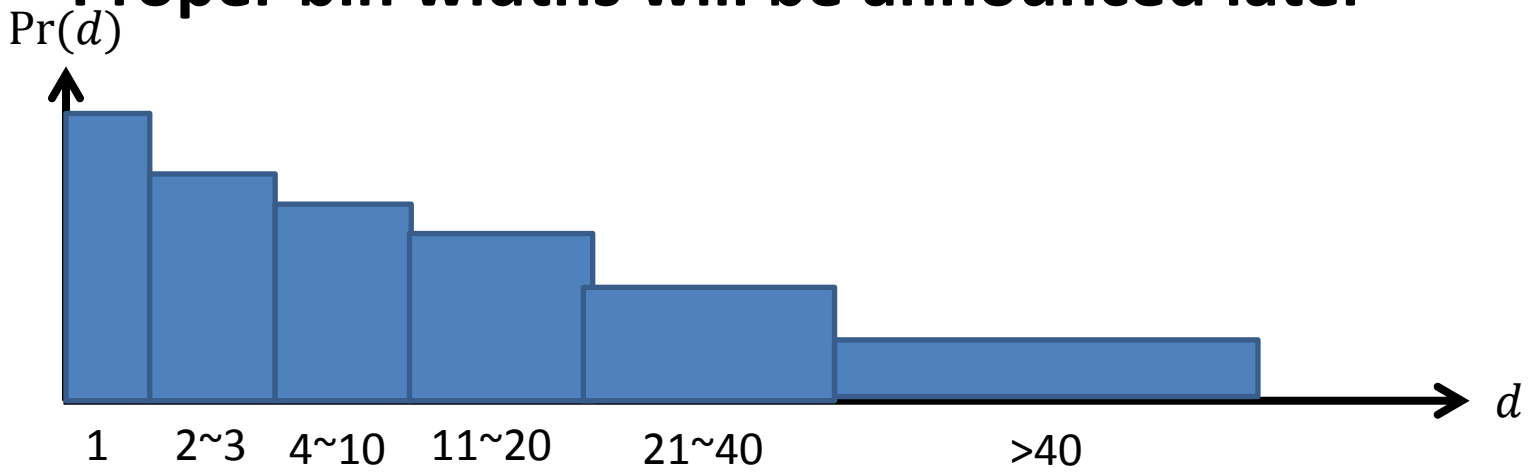
- Starting from Nov 21st
  - You will be given the a fully observed public graph
  - The query system for the public graph is released for you to use
  - The number of queries is not limited
- 36 hours before the deadline (i.e. on Dec 9th 10:00am)
  - A private graph will be released, with a seed subgraph of n_s observed nodes
  - **Quota**: Each team is allowed to send at most n_q queries (n_q & n_s will be announced later)
  - **Your score will be graded using the private data**

Evaluation

- Your want to use the sampled graph H to estimate the following properties of the original graph G
  - Degree distribution
  - Top 100 nodes with the highest closeness centrality values
  - Distributions of each node attribute and each edge attribute

Degree Distribution (1 / 2)

- Probability distribution
  - X-axis: Degree d
  - Y-axis: Probability $\Pr(d)$, fraction of nodes of degree d
- Bin combination before evaluation
  - **Proper bin widths will be announced later**



Degree Distribution (2 / 2)

- Evaluation metric: KL divergence
  - $D_{KL}(P, Q) = \frac{1}{2}(D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P))$
  - $D_{KL}(P \parallel Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}$
  - P: True probability distribution over degrees in G
  - Q: The probability distribution over degrees in H
  - The smaller, the better

Closeness Centrality

- Formula
  - $CC(v) = \frac{n-1}{\sum_{u \in V, u \neq v} d(v,u)}$
  - $d(v,u)$: Shortest path length from v to u
- Evaluation metric: Average true rank
  - $R = \frac{1}{k} \sum_{i=1}^{k} rel_i$
  - We evaluate top-k ($k = 100$) nodes in H
  - i : The rank of a node v in the sampled graph H
  - rel_i : The true rank of v in the original graph G
  - The smaller, the better

Attribute Distributions

- For each attribute, there is a probability distribution
  - Probability , fraction of nodes or edges of some attribute value
- The attribute values are discrete (e.g. male/female)
- Evaluation metric: KL divergence

Homework Files

- *HW3/*
  - *query.py* # Connection to the query system
  - *report.docx* # Report format description
- Public graph inside the query system
  - Download links
    - http://140.112.31.186/SNA2014/hw3/public_nodes.zip
    - http://140.112.31.186/SNA2014/hw3/public_edges.zip
    - http://140.112.31.186/SNA2014/hw3/public_readme.txt

Submission Files

- *hw3_team_{id}/* # e.g. “hw3_team_1”
  - *report.pdf* # Report
  - *sample.txt* # Your sampled graph for the private graph
  - *degree.txt* # Degree distribution
  - *closeness.txt* # Top 100 nodes
  - *node_attr_1* # Distribution of node attribute 1
  - …
  - *node_attr_k_V* # Distribution of node attribute k_V
  - edge_attr_1 # Distribution of edge attribute 1
  - *Makefile*
  - *Your code*

query.py

- An example of connection to the query system
- Inputs, outputs are the same as the system
- Python 3 code
- You can write your own connection program

```
python3 query.py team [node]
```

sample.txt

```
2014 505                    # A line records an edge
2014 2222
2222 9908
5032450 45345
45043 432580
…
67 3245
67 11654
```

m edges

degree.txt

```
1 1 0.64        # 64 % of nodes with degree 1
2 3 0.3         # 30 % of nodes with degree 2 ~ 3
4 10 0.04
11 20 0.015
21 40 0.003
41 0 0.002      # 0.2 % of nodes with degree ≥ 41
                # (the 2nd integer is filled 0)
```

closeness.txt

```
3466    # Top 1 node with the highest closeness centrality values
715     # Top 2
61
65767
...
465     # Top 100
```

node/edge_attr_k.txt

- Values are integers that should be in ascending order

```
2 0.33    # 33 % of nodes/edges with attribute value = 2
3 0.02    # 2 % of nodes/edges with attribute value = 3
5 0.3
7 0.01
11 0.34
```

Report

- Content
  - Performance of your experiments
  - Description of your models
- Format
  - Both English and Chinese are welcomed
  - No more than 8 pages
  - Please follow report.docx, filling in the table
  - Convert your report to the PDF file

Running Your Program

- Environment: NTU CSIE workstation
- TA will type the following command to test your code

```
make                    # run code in Makefile
```

- Your program should automatically output *sample.txt* in the same directory
- If you need to use special packages / libraries (except Networkx), please include them in your submissions
- If there is difficulty in running your code (e.g. cannot run with Makefile, packages cannot be included), please remind us how to correctly run your program in the report

Tips and Hints

- Your sampling models should be general
  - The private network might not be similar to the public network
  - Take other heterogeneous graphs as your training data
- You can use any means to estimate the distributions. They don’t need to be identical to the distribution of your sampled graph.
- You are encouraged to implement sampling models from papers or slides in this homework

Be Careful

- **NEVER** take time to seek or guess the private graph
  - The information (e.g. number of nodes / edges) will be modified if necessary
- **NEVER** try to attack the query system (e.g. denial of service)
  - The system records every query in the database
  - **Your might fail this course such behavior is identified**

Grading

- Evaluation 70%
  - In terms of relative performance of all teams
- Report 30%
  - Format
  - Experiments
  - Models

Submission

- **Deadline: 2014/12/10 22:00**
- Compress the directory *hw3_team_{id}/* to a **ZIP** file
- Please contact TA by email, CEIBA board or TA hour if you have any questions about the homework

To Be Announced

- The parameters for private data will be announced next week
  - Number of seed nodes n_s for the private data
  - Maximum number of queries n_q for the private data
  - Proper bin widths of the degree distribution