Assignment 2 Solution

1 Problem 1 (20%)

In this question, we consider the problem related to **rich-get-richer** model. Suppose there is a graph with N nodes, and the link creation between these nodes is based on the following rules:

- Nodes are created one by one from 1 to N. When the node j is created, it has one outbound link to one of its previous nodes.
- When the node j is created, the probability of creating a link to the node i (which is chosen uniformly at random from all earlier nodes) is p.
- With probability $1 - p$, node j creates a link to the node that i links to.

Q1: (20%) Please write the following 3 expressions: (You should write expressions with j, t, p and k)

- the in-links of node j as $x_j(t)$ at a time t, $(t \geq j)$.
- the number of nodes with at least k in-links as $F(k)$.
- the number of nodes with exact k in-links as $f(k)$.

2 Problem 2 (30%)

In this question, we consider the model of **Cascading Behavior**. Suppose everyone uses behavior A in the following network(Fig. 1) initially, and then a new behavior B is introduced. q_B is the threshold of this behavior B, which means each node would switch behavior to B if at least q_B fraction of its neighbors adopt behavior B. Please answer the following questions.

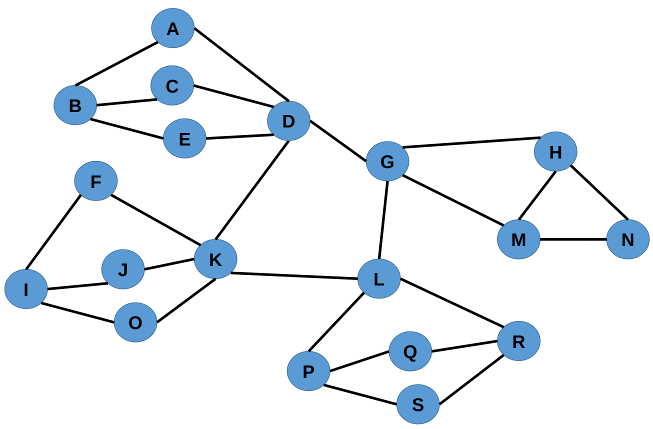

Figure 1: Graph of Problem 2.

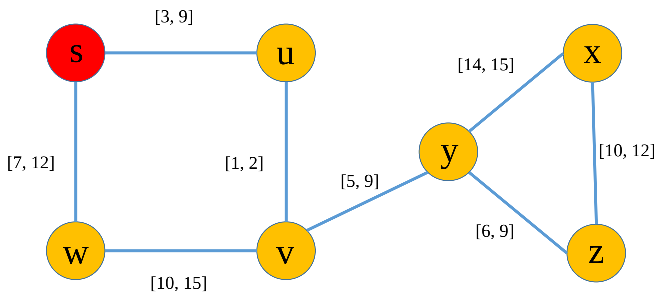

Figure 2: Graph of Problem 3.

Q1: (5%) Find four clusters in the network (Fig. 1) whose density is higher than $\frac{1}{2}$, and with the property that no node belongs to more than one of these clusters.

Q2: (10%) Suppose $q_B = \frac{1}{3}$, find the minimum number of nodes that can act as initial adopters of behavior B and finally all nodes would adopt B. i.e., find the minimum set of nodes which can cause a complete cascade of adoption of B.

Q3: (5%) Suppose you can only choose just one node to adopt behavior B. If we want to achieve the cascade of B complete, please determine the maximum value of q_B so that we can achieve this target.

Q4: (10%) Given x as the number of initial adopters of behavior B, we set $S(x)$ as the set of possible values for q_B under which we can cause a complete cascade finally. We define function $f(x)$ as follows:

$$f(x) = max\{q|q \in S(x)\} \tag{1}$$

Please draw the figure of $f(x)$ with discrete values ($x = 1, 2, 3, ...$) and connect those discrete dots. Describe the shape of that curve.

3 Problem 3 (20%)

In this question, we study the spread of a disease among a set of people. In the following figure 5, each node represents a person. The edges between different people denote the contact between them. There is a time interval marked on each edge showing the period of contact occurred (e.g. [2, 4] means these two persons contact between time 2 to time 4). And we just observe the spread of disease from time 0 to time 15. Please answer the following questions.

Q1: (10%) Suppose the node "s" is the only individual who had the disease at time 0. Which nodes could potentially have acquired the disease by the end of the observation period, i.e. the time 15?

Q2: (10%) Suppose all nodes have this disease at time 15. And we assure that this disease has not been introduced from other sources. Can you find a single number in these time intervals to change (i.e. designating the start or end of one of these time intervals), so that in the resulting network, it's possible for the disease to spread from "s" to every other node?

4 Problem 4 (30%)

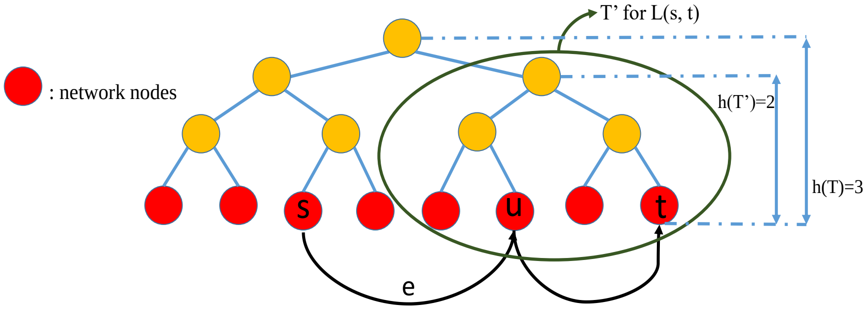

Figure 3: Graph of Problem 4.

In this question, we study the decentralized search in a balanced tree hierarchy (a small world). In the following figure 6, there is a balanced b-ary tree representing the hierarchical organization of CUHK. The root represents CUHK, the second level's nodes denote different collages, the third level's nodes denote organizations under collages and the leaves denote the students. If one student want to contact a stranger, he may try to find a friend who is "closest" to this stranger in the tree.

Especially, we first define the notion of "tree distance": given two leaf nodes "v" and "w", $L(v, w)$ denotes the subtree T which is rooted at the lowest common ancestor of v and w. And $h(v, w)$ denotes the height of this subtree. For example, in figure 6, L(u, t) in the circle is the subtree for node "u" and "t" and h(u, t)=2. We call h(u, t) as the "distance" between these two leaf nodes u and t.

Q1: (10%) We can generate a random network on these leaf nodes. In the generated network, a node is more likely to know "close" nodes than "distant" nodes according to the hierarchy of tree. To achieve this, we can define a probability distribution of node v creating an edge to any other node u as: $p_v(u) = \frac{1}{Z} b^{-h(v,u)}$ where $Z = \sum_{u \neq v} b^{-h(v,u)}$ is a normalizing constant.

Please show that $Z \leq log_b N$. (Hint: for a balanced b-ary tree T, its height is $h(T) = log_b N$, where N is the number of leaves in the tree. You can compute Z by considering h(v, u) with all possible values).

Q2: (10%) For two leaf nodes v and t in the network, we define T' as the subtree of $L(v, t)$ such that: 1)the height of T' is $h(v, t) - 1$; 2) T' contains t; 3) T' does not contain v. For example, in figure 6, T' of $L(s, t)$ is the tree in the circle.

As stated, there can be an edge e from v to a random node u with probability sampled from p_v. We can say that e points to T' if u is a leaf node in T'. Please show that the probability of e pointing to T' is no less than $\frac{1}{blog_b N}$ if we define $h(v, u) = d$ and u is a leaf node in T'. (Hint: use the result in Q1).

Q3: (10%) In fact, we can ensure that every node v in the network to have exactly k outgoing edges: For each node v, we repeatedly sample a random node w according to p_v and create edge (v, w) in the network. We continue this until v has exactly k neighbors. k is a parameter which can be controlled by set $k = c(log_b N)^2$, where c and b are constants.

Please show that when N grows very large, the probability of v not having any edge pointing to T' is no more than $N^{-\theta}$, where θ is the positive constant which you need to compute. For each node, you can assume that each of its k outgoing edges is independently created. (Hint: use the results in Q2 and use the equation $lim_{x \to \infty}(1 - \frac{1}{x})^x = \frac{1}{e}$.)

5 Problem 1 (20%)

Q1: (20%) 1. Solve $X_j(t)$

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t} \tag{2}$$

$$X_j(t) = \frac{p}{q}((\frac{t}{j})^q - 1) = \frac{p}{1-p}((\frac{t}{j})^{1-p} - 1) \tag{3}$$

2. Solve $F(k)$

$$F(k) = t(\frac{q}{p} \times k + 1)^{-\frac{1}{q}} = t(\frac{1-p}{p} \times k + 1)^{-\frac{1}{1-p}} \tag{4}$$

3. Solve $f(k)$

$$f(k) = -\frac{dF}{dk} \tag{5}$$

$$f(k) = \frac{t}{p}(1 + \frac{q}{p} \times k)^{-(1+\frac{1}{q})} = \frac{t}{p}(1 + \frac{1-p}{p} \times k)^{-(1+\frac{1}{1-p})} \tag{6}$$

6 Problem 2 (30%)

Q1: (5%) Cluster 1: {A, B, C, D, E}
Cluster 2: {F, I, J, K, O}
Cluster 3: {G, H, M, N}
Cluster 4: {L, P, Q, R, S}

Q2: (10%) the minimum number of nodes to trigger cascade: 2, e.g., {D, L}

Q3: (5%) The maximum of q_B so that it can achieve cascade is 0.25, e.g., D

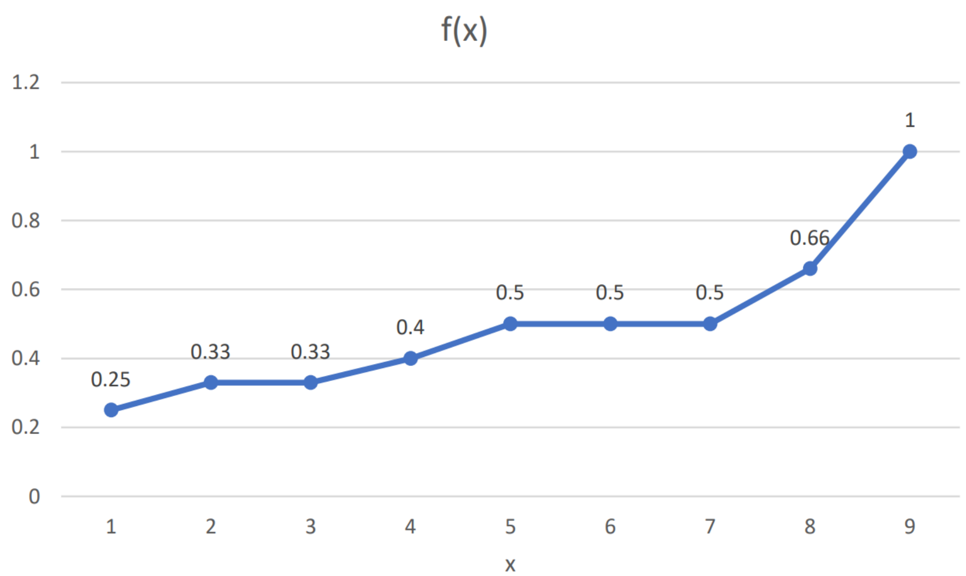

Figure 4: Graph of Q4.

Q4: (10%)

7 Problem 3 (20%)

Q1: (10%) node “u”, “w” and “v”.

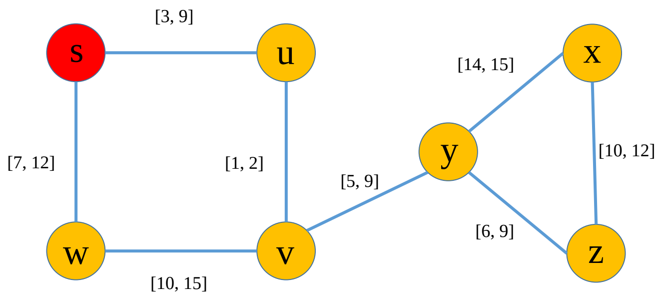

Figure 5: Graph of Problem 3.

Q2: (10%) The answer is not unique. For example, we can change the time interval on edge "u-v" to $[1, \{n|n \geq 5\}]$.

8 Problem 4 Solution(30%)

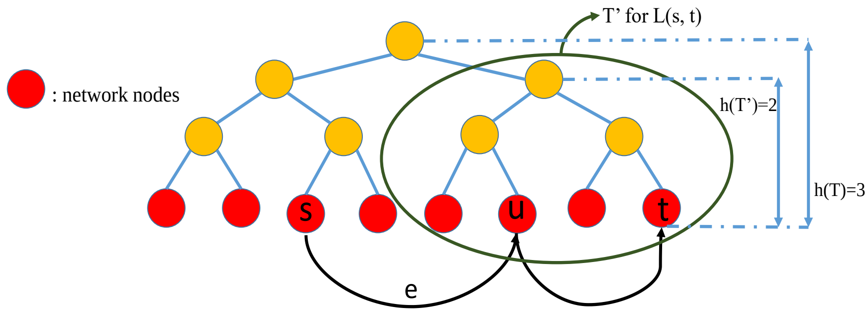

Figure 6: Graph of Problem 4.

Q1: (10%) For a balanced b-ary tree T, its height is $h(T) = log_b N$, where N is the number of leaves in the tree. In similar way, the number of leaves of a tree of height h(T) is $b^{h(T)}$.

Let $T_{v,d}$ be the subtree of T containing v and of height d. The leaves u of $T_{v,d}$ - $T_{v,d-1}$ satisfy $h(v,u)$ = d. **Thus, there are $b^d - b^{d-1}$ nodes satisfying $h(v,u)$ = d.**

Notice that we can group nodes "u" according to $h(v,u)$, which ranges from 1 to logN. And we now have know that the number of nodes "u" satisfying $h(v,u) = i$ is $b^i - b^{i-1}$. This substitution allows us to write that:

$$Z = \sum_{u \neq v} b^{-h(v,u)} = \sum_{i=1}^{log_b N} (b^i - b^{i-1}) b^{-h(v,u)} = \sum_{i=1}^{log_b N} (b^i - b^{i-1}) b^{-i} = \sum_{i=1}^{log_b N} (1 - \frac{1}{b}) \leq log_b N \tag{7}$$

Q2: (10%) We can assume that $d = h(v,t)$, and we can see that there are b^{d-1} leaves in T'. We now know that $h(v,u) = d$, and the probability of "v" linking an edge to "u" is $p_v = \frac{b^{-d}}{Z}$. Therefore, the probability of "v" has an edge pointing to T' is no loss than (we have know from Q1 that $Z \leq log_b N$):

$$b^{d-1} p_v = b^{d-1} \frac{b^{-d}}{Z} = \frac{1}{bZ} \geq \frac{1}{blog_b N} \tag{8}$$

Q3: (10%) We can first assume that the creation of edges from "v" to be independent. From Q2, we have know that the probability of 'v" has an edge pointing to T' is larger than $\frac{1}{blog_b N}$, and the probability of not having an edge pointing to T' is thus no more than $1 - \frac{1}{blog_b N}$. Therefore, the probability of v having no edges to T' is smaller than:

$$(1 - \frac{1}{blog_b N})^k = (1 - \frac{1}{blog_b N})^{c(log_b N)^2} = (1 - \frac{1}{blog_b N})^{b(log_b N) \frac{c}{b} (log_b N)} \tag{9}$$

when N grows very large, $(1 - \frac{1}{blog_b N})^{b(log_b^N)} = \frac{1}{e}$, and we can write equation 9 as:

$$(1 - \frac{1}{blog_b N})^k = (1 - \frac{1}{blog_b N})^{b(log_b N) \frac{c}{b} (log_b N)} = (\frac{1}{e})^{\frac{c}{b} \frac{In(N)}{In(b)}} = N^{-\frac{c}{bIn(b)}} \tag{10}$$

Thus, $\theta = \frac{c}{bIn(b)}$.