UNIVERSITY OF EDINBURGH, SCHOOL OF INFORMATICS

Distributed Systems Assignment

Andres Monteoliva s1782267

April 1, 2018

1 THEORETICAL ANALYSIS OF RUMOUR SPREADING

Note: *Given that the question 4 is just a particular case of the more general question 5, in this document I would present my results for both question together.*

The spreading of rumours on networks is a well-studied problem in academia [1]. The purpose of this question is to study the behavior of the PUSH protocol in random graphs.

For a fully-connected graph of n nodes G_n (i.e. a complete graph) we have that every node will get the message with high probability in:

$$T(n) \sim log(n) + ln(n) + o(log(n)) \tag{1.1}$$

With high probability (w.h.p), an standard term in probabilistic algorithms theory, will be defined in this assignment such that, in a graph with n vertex (with n a big enough number) the probability that the gossip algorithm will converge in a number of rounds is:

$$1 - 1/n \tag{1.2}$$

so for big enough n, we will have that w.h.p. can reformulate as:

$$1 - o(1) \tag{1.3}$$

1.1 MODEL OF OUR RANDOM GRAPH

As we know the result for the termination time of the algorithm in a complete graph, this will serve as the starting point to the study of the termination time in random graphs.

As random graph we will follow the model of random graph proposed by Erdos-Renyi [2], with a graph G(n,p), being n the number of nodes and p the probability that a pair of nodes is connected (e.g. that an edge exists). We must be aware that our model represents a random evolving graph, so in each round we will have a different graph that fulfill these condition. So for any set of rounds t+1, the graph underlying the network can be understand as the sequence:

$$G_1(n,p), G_2(n,p), ..., G_t(n,p), G_{t+1}(n,p) \tag{1.4}$$

For a more detailed information of the graph model, please refer to the Assignment page.

1.2 Question 4. Analysis of termination time with p constanst

The analysis of the termination time will follow the same approach of the analysis for the complete graph, as shown in the slides of the course. Let I_t be the set of informed nodes in the end of round t, and let U_t be the set of non-informed nodes in the end of round t. The analysis will be divided in three phases, namely:

1. $1 \leq |I_t| \leq (n-1)/log(n) \longrightarrow$ (number of nodes informed is small)
2. $(n-1)/log(n) \leq |I_t| \leq n - n/log(n) \longrightarrow$ (~ half of nodes is informed)
3. $n - n/log(n) \leq |I_t \leq n \longrightarrow$ (number of nodes uninformed is small)

To calculate the number of rounds needed for all the nodes becoming aware, following the rational of the lecture's slides, we will calculate the increase of number of informed nodes in each round (i.e. nodes that are added to the set of informed nodes in round t , I_t). So we are going to calculate the expected value of nodes in I_{t+1} that were not informed in round t :

$$\mathbb{E}[|I_{t+1} \setminus I_t|] = \sum_{u \notin I_t} P[u \in I_{t+1}] \tag{1.5}$$

The probability of being in the informed set is complementary to the probability of still being in the uninformed set in round t+1:

$$\sum_{u \notin I_t} P[u \in I_{t+1}] = \sum_{u \notin I_t} 1 - P[u \in U_{t+1}] \tag{1.6}$$

Let's now focus on the probability that one informed node u is NOT informed by an informed node v in a round t. In our **random graph setting**, this probability can be presented as:

$$P(connected) * P(NOTreceivingthemessage|connected) + P(NOTconnected) \tag{1.7}$$

As a reminder, the formal definition for being "connected" is that the edge between node u and node v is present in the given round. In our case, the expected value that this edge exists would be: $P(connected) = p = 0.3$. The above expression can be reformulated as follows:

$$P[u \in U_{t+1}] = p * P(\neg msg|connected) + (1-p) \tag{1.8}$$

The probability of receiving a message from node v given that the connection exists, will be the the inverse of the degree of node v in that determined round. The expected value of the degree of node v would be:

$$\mathbb{E}(deg(v)) =< deg(v) >= p * (n-1) \tag{1.9}$$

Therefore, we can rewrite equation 1.7 as:

$$P[u \in U_{t+1}] = p * (1 - \frac{1}{p * (n-1)}) + (1-p) \tag{1.10}$$

It should be noted that 1.10 is just valid for p being a constant, as the present case of question 4, where p=0.3. Later, it will be discussed in question 5 when p(n). Equation 1.10 can be rewritten as:

$$P[u \in U_{t+1}] = p * (1 - \frac{1}{p * (n-1)}) + (1-p) = 1 - \frac{1}{n-1} \tag{1.11}$$

Now, we want to generalise this probability to the set of all informed nodes. Assuming that each informed node will choose a random neighbor blindly and that the existence of an edge in the graph is not dependent of past existence of that proper edge we can conclude that the probability of u receiving a message from one of the nodes in I_t will be:

$$P_{v_0} \cap P_{v_1} \cap P_{v_n} ... \simeq \prod_{v \in I_t} P(v \rightarrow u) = (1 - \frac{1}{n-1})^{|I_t|} \tag{1.12}$$

being $|I_t|$ the cardinality of the set of informed nodes.

So following equation 1.5 and 1.6, and generalizing the result for the whole set of uninformed nodes we have:

$$\mathbb{E}[|I_{t+1} \setminus I_t|] = \sum_{u \notin I_t} P[u \in I_{t+1}] = \sum_{u \notin I_t} 1 - P[u \in U_{t+1}] = \sum_{u \notin I_t} 1 - (1 - \frac{1}{n-1})^{|I_t|} \tag{1.13}$$

As we can recall, this exactly the same expression as the one shown by the lecture slides of the case of the complete graph. Therefore, we apply the same rationale for the three phases, and following the steps of the aforementioned proof on the complete graph case. Following the same approach for the three different phases mentioned above, we demonstrate that the convergence number of rounds for a constant factor $p = 0.3$ will converge w.h.p in the following number of rounds:

$$T(n) \sim log(n) + ln(n) + o(log(n)) \tag{1.14}$$

In conclusion, we have proven (building on top of the proof for the complete graph) that the **termination time (in number of rounds) will not be affected asymptotically if we introduce a constant probability** $p = 0.3$**.** What this result says is that the PUSH protocol convergence time is **robust** against variations on the density of the underlying graph.
However, it should be mentioned that we have done several assumptions, which may alter our result in case they were proven invalid.

Assumptions

1. We are assuming that $p = 0.3$. As it will be shown in the question 5, if p is very small the underlying graph G will become very sparse. Therefore, assumptions like the expected value of the degree of a node (as in 1.9 and 1.10) cannot be formulated.

2. In step 1.12, we are assuming that the probability that an uninformed node u will not receive a message from any of the nodes within the set of informed nodes can be rewritten as the product of the probabilities for each node.

3. To follow the same steps as the proof of the complete graph shown in class, we are assuming that the inductive step to calculate $\mathbb{E}[|I_{t+1} \setminus I_t|]$ is valid in the random graph model. The rationale for this assumption is that the sequence of random evolving graphs depicted in equation 1.4, for a big number of rounds t and a big number of vertex n can be viewed asymptotically as a constant graph with density d:

$$d = p * n \tag{1.15}$$

1.3 Question 5. General analysis of termination time for p(n)

The reasoning and results of this section follow the paper of Fountoulakis et al. in *Reliable Broadcasting in Random Networks and the Effect of Density* [4].

In this section we need to reason how the number of rounds needed for the termination of the PUSH protocol is going to be dependent on p(n). In this case, p is not a constant as in question 4 but an arbitrary function dependent on the number of nodes on the graph. Theorem I.1 in the aforementioned research paper tell us that if

$$p \geq \frac{\alpha(n) ln(n)}{n} \tag{1.16}$$

with $0 \leq \alpha(n) \leq ln^{1/9}(n)$ and $lim_{n\to\infty}\alpha(n) = \infty$ then:

$$T(n) \sim log(n) + ln(n) + o(log(n)) \tag{1.17}$$

Note: Actually, in this research paper they use a tighter bound, but for the scope and purpose of this assignment, we would keep using the bound studied in class.

What theorem I.1 [4] states is that the number of rounds needed for the PUSH protocol to terminate with high probability is going to be the same as the complete graph case if and only if the equation 1.16 is satisfied. However, below this threshold, our analysis approach (the one taken in question 4) is not valid anymore. Given that now we are studying random graphs with p(n), the properties of them will vary. In particular, in question 4 we assumed that the expected value of the degree of each of the nodes in our graph would be $deg(v_i) \sim p * (n-1)$.

More formally, being $S \subset V$, with V the set of informed nodes, the expected neighbors of any $v \in V \setminus S$ in S would be $p|S|$. Lemmas IV.1 and IV.2 in [4] prove that with $p \geq \frac{\alpha(n) ln(n)}{n}$ then,

all vertex in a random graph $G(n,p)$ with a big enough S will have the right degree in S.

If we proceed as in question 4, where our approach was to calculate the rate of growth of informed nodes in each round t, namely, calculating the expected value:

$$\mathbb{E}[|I_{t+1} \setminus I_t|] = \sum_{u \notin I_t} P[u \in I_{t+1}] \tag{1.18}$$

then, taking an **inductive step** through each round, prior havind divided the protocol in 3 phases, we face a major **problem**.

In that case, we assumed that the degree of each node would be equal to its expected value, exploiting the fact that all vertex would have the "right degree" if equation 1.16 holded.

However, if this equation does not hold (in very sparse graphs), the degree of each vertex deviates from its expected value. This becomes crucial in the last phase of the protocol, where the numbers of rounds needed for termination will increase considerably in very sparse graphs. To make this clearer, below is shown a figure from [4], where the number of stages needed by the protocol needed to terminate is plotted against the number of edges. We can see that when the density of the graphs is very small ($p << 1$) the number of rounds increases:

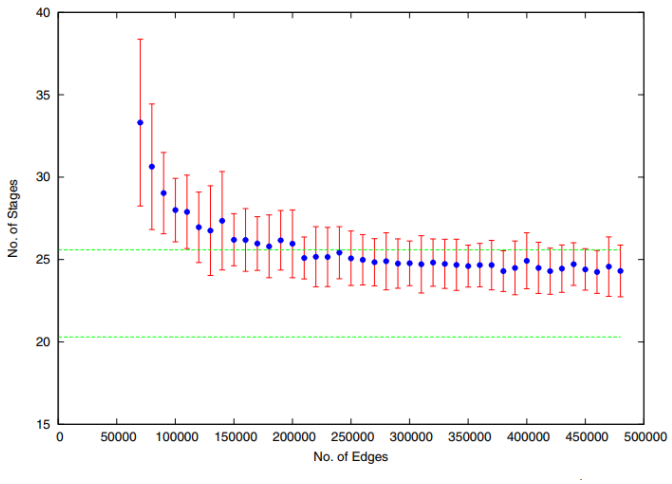

Figure 1.1: Simulation for sparse graphs, with $n = 10^4$ [4]

In conclusion, in question 5, relying in the mathematical proofs of [4], we have shown that **the PUSH protocol, when eq.1.16 does not hold, deviates from the standard termination time calculated in question 4 (increases when p tends to 0). An inductive analysis is not valid in this case, given that for very sparse graphs, the degree of the nodes in the last phases of the protocol is not equal to its expected value.**

2 REFERENCES

1 PITTEL,B. (1987) On spreading a rumor. In *SIAM Journal on Applied Mathematics,* 47(1), 213-223.

2 ERDOS,P. et RENYI, A.(1959) On random graphs. In *I. Publicationes Mathematicae (Debrecen)* , 6, 290-297.

3 CRISOSTOMO,S. et al. (2009) Analysis of probabilistic flooding: How do we choose the right coin?. In *Communications, 2009. ICC'09. IEEE International Conference on (pp. 1-6)*. IEEE.

4 FOUNTOULAKIS,N. et al. (2010) Reliable broadcasting in random networks and the effect of density. In *INFOCOM, 2010 Proceedings IEEE (pp. 1-9)*. IEEE.