Gossip

Assume that we have a large, well-connected network of $n$ servers. The problem that we are investigating in this setting is the natural problem of finding optimal strategies to disseminate a message from a server $s$ to the rest of the network. Furthermore, we assume that each link might fail with the probability of $p$. Note that in this setting, nodes are not aware of the source of information beforehand. Here are a couple of intuitive ideas to solve this problem:

A simple solution for this problem would be flooding. In this strategy, once a node receives the new message for the first time, forwards it to all its neighbors. This idea will clearly convey the message to all servers. However, each server will receive the message multiple times and the communication has a huge amount of redundancy.

Another possible strategy would be passing the message to the neighbors on a predefined spanning tree $T$. In this strategy the message dissemination will be finished in a time proportional to the tree length and there will be an overall number of $n-1$ message transfers. However, this strategy clearly lacks robustness, i.e., once a link fails in this protocol, some nodes will never get the new message.

In order to find a solution for this problem which is both optimal and robust, we use a randomized idea. In this very simple idea, each node that knows the new message, selects one of its neighbors uniformly at random in each round and passes the information to it. Now, let’s address the following natural questions that arise about this strategy’s performance.

For the sake of simplicity, from now on we assume that $p = 0$ and the graph is complete.

Q1. What is the expected number of messages that a specific node receives in each round?

Assume that servers $i_1, i_2, \ldots, i_t$ are aware of the new message. We define $X_{ij}$ random variables as indicators of whether $i_j$ sends the new information to the specified node in this round or not. The linearity of expectation gives that this expectation is less than or equal to $1$.

$$
E[\# \text{ of received Messages}] = E\left[ \sum_{j=1}^{t} X_{ij} \right] = \sum_{j=1}^{t} E[X_{ij}] = t \times \frac{1}{n} = \frac{t}{n} \leq 1
$$

Q2. Are there any high probability results for the expected number of messages that a specific node receives in each round?

Chernoff bound gives us:

$$
Pr\{X > (1 + \delta)\mu\} < e^{-\frac{\delta^2 \mu}{3}}
$$

However, in this case as the expectation of $X$ is less than one and $\delta$ has to be a constant, we cannot obtain high probability results from chernoff bound. One idea that works here is to
add independent random variables to \( X = \sum_{j=1}^{t} X_{ij} \) to obtain new random variable \( Y \) with mean \( \log n \). Then:

\[
\Pr\{X > \log n\} < \Pr\{Y > \log n\} < e^{-\log n / 3} = \frac{1}{n^3}.
\]

**Q3.** What is the expected number of messages that a specific node receives in \( \log n \) rounds?

By linearity of expectation, the answer is clearly \( \log n \times \mathbb{E}[X] < \log n \).

**Q4.** Are there any high probability results for the expected number of messages that a specific node receives in \( \log n \) rounds?

We can write the total number of messages that a specific node receives in \( \log n \) rounds as summation of \( n \log n \) indicator variables \( X_{i,r} \) indicating if a message has been received from node \( i \) in round \( r \). As mean is less than or equal to \( \log n \) we can use chernoff bound or the technique applied in Q2 to show that the number of messages that a specific node receives in \( \log n \) rounds is less that \( 2 \log n \) with high probability.

**Q5.** What is the maximum number of messages received by any node in one round with high probability?

As seen in Q2, the probability of a specific node receiving more than \( \log n \) messages in one round is negligible. Using union bound, the probability of the maximum number of messages received by any node exceeding \( \log n \) will be at most \( n \) times higher. Thus, maximum number of messages received by any node in one round will be with high probability less than \( \log n \).

**Q6.** What is the expected maximum number of messages received by any node in one round?

Using chernoff bound and the technique used in Q2, the probability of a server receiving more than \( X > 1 \) messages in one round can be bounded by \( e^{-cX} \). Let \( X \) be \( \frac{\log n}{c'} \). Then, the probability of one node receiving more than \( \frac{\log n}{c} \) messages will be at most \( \frac{1}{n^{c'}} \). Thus, the probability of maximum number of messages received by any node being more than \( X \) will be at most:

\[
1 - \left( 1 - \frac{1}{n^{c'}} \right)^n = 1 - e^{-n^{c-c'}}
\]

One can easily see by increasing \( c' \) to a proper constant, this probability increases very sharply. This shows that the maximum number of messages received by any node is greater than a constant factor of \( \log n \) with high probability. So, expected maximum number of messages received by any node a round is \( \Theta(\log n) \).

**Q7.** What is the maximum number of messages received by any node over \( \log n \) rounds with high probability?

Since maximum number of messages received by any node in one round is less than \( \log n \) with high probability, using union bound, this fact remains true for maximum number of messages received by any node over \( \log n \) round with high probability.
Q8. What is the expected maximum number of messages received by any node during $\log n$ rounds?

An argument similar to Q6 works here to show that expected maximum number of messages received by any node during $\log n$ rounds is $\Theta(\log n)$.

Now, we get back to the analysis of the gossip protocol. Assume that $X_i < \frac{n}{2}$ nodes are informed about the new message at the end of round $i$. The probability of the event that a specific uninformed node receives the message is:

$$P = 1 - \left(1 - \frac{1}{n}\right)^{X_i} = 1 - \left(1 - \frac{X_i}{n} + \left(\frac{X_i}{n}\right)^2 - \cdots\right) > \frac{X_i}{n} + \left(\frac{X_i}{n}\right)^2 > \frac{X_i}{4n}$$

Hence, expected number of newly informed nodes in next round will be:

$$\mathbb{E}[\text{Number of newly informed nodes}] > \frac{(n - X_i)X_i}{4n} > \frac{\frac{n}{2}X_i}{4n} = \frac{X_i}{8}$$

Then:

$$\mathbb{E}[X_{i+1}] \geq \mathbb{E}[X_i] \times \left(1 + \frac{1}{8}\right)$$

So:

$$\mathbb{E}[X_i] \geq \left(1 + \frac{1}{8}\right)^i$$

If we repeat the algorithm for $c \log n$ rounds, we will have

$$\mathbb{E}[\text{Number of informed nodes}] \geq \left(1 + \frac{1}{8}\right)^{c \log n} = n^{c'}$$

Using Markov inequality, it can be easily shown that the probability of the number of informed nodes being less than $\frac{n}{2}$ at the end of $c \log n$th round will be at most $n^{-c'}$. So far, we have shown that in $O(\log n)$ rounds more than half of nodes will be informed with high probability. In the next note, we will show that by changing the transfer strategy from push (i.e. informed nodes sending information) to pull (i.e. uninformed nodes asking for information) at this point, an optimal gossip protocol will be obtained.