Assignment 3 Report
s3737548 Yuxin Ye

Q1.
Compare to GreedySel, EnumSel can produce a much more accurate result. Its approximation guarantee (1-1/e) is more accurate than GreedySel (1/2(1-1/e)).
Compare to EnumSel, PartSel is much faster. By partitioning the billboard data, it has a smaller datasets value (Cm), that is much smaller than the entire billboard datasets (U), thus it's time complexity O(mL |T||Cm|^5) is smaller than EnumSel O(|T||U|^5).
PartSel can also produce good approximation guarantee in comparison with GreedySel. If the values of Θ and m are small, its guarantee ($\frac{1}{2}^{ceiling[\log_{\left(1+\frac{1}{\Theta}\right)} m]}$ $(1 - 1/e)$) can reach (1-1/e).

Q2.
The Naïve greedy method cannot achieve any theoretical guarantee. Because it cannot achieve optimal result for a billboard set. For example, give two billboards (a) and (b).
(a) with influence of 1, cost 1, marginal influence (1)
(b) with influence of x, cost x + 1, marginal influence(x/(x+1))
Budget is x+1.
The naïve greedy would pick board (a) with influence 1, as its marginal influence is greater than (b). But it cannot pick board (b) again because its budget has run out.
However, the optimal solution should be (b) with influence x, while x >= 1.
Because influence x can be arbitrarily large, so we aren't able to give a theoretical guarantee of this method.

Q3.
We cannot compute pr(S, t_j) as the aggregation of pr(b_i, t_j).
Because we need to avoid influence overlap. For example, when one person sees the same billboard in different places in a single trajectory, they are only considered been influenced once.
We need to use (1 - pr(b_i, t_j)) to find the probability that one trajectory is not influenced by a single billboard. Use a product to enumerate all the billboards, find the probability that all trajectories are not influenced by these billboards. In another words, find the probability of people who are not meeting these billboards.
Then we need to use 1 minus this entire product, to get the probability of trajectories that are going to be influenced by at least one of the billboards. In another words, the probability of people who are influenced by at least one of the billboards.
The equation from the research paper is given to represent this process:

$$pr(S, t_j) = 1 - \prod\nolimits_{b_i \in S} (1 - pr(b_i, t_j))$$

Q4.
From the research paper, the value of Θ is used as a threshold to control the maximum possible overlap ratio between the pairs of clusters. First each billboard is initiated as its' own cluster, then if two cluster's overlap ratio are larger than Θ, we iteratively merge two clusters into one. Finally obtain an approximate Θ-partition when no clusters in all the billboard dataset can be merged.

When Θ is small, the influence overlap ration between two clusters will be small, the final result will have optimal clusters with smaller overlap ratios between them.
When Θ is large, the overlap ratio between two clusters will be larger. Therefore, the result will have more overlapped clusters.

If quality is required, Θ can be set small to produce higher quality clusters. However, it will take more time to operate the algorithm.
If efficiency is required, Θ can be set to a larger value to get faster but less accurate result.