Evaluation of Seed Selection Strategies for Vehicle to Vehicle Epidemic Information Dissemination

Richard Kershaw and Bhaskar Krishnamachari
Ming Hsieh Department of Electrical Engineering,
Viterbi School of Engineering
University of Southern California, Los Angeles, CA 90089
{rkershaw, bkrishna}@usc.edu

***Abstract*—We consider the problem of how to identify a set of seed nodes in order to disseminate content efficiently in a vehicular network. We consider several relevant dimensions including proximity, encounters, speed, etc. and identify and taxonomize a number of candidate strategies. We comparatively evaluate these strategies using a set of real vehicular traces (Taxis in Beijing). We conclude that identifying seeds based on their speed, while eliminating redundant and isolated nodes, is the most effective approach, performing significantly better than the previously random seed strategy.**

I. Introduction

With the recent announcement by the U.S. Dept. of Transportation that it plans to ask car manufacturers to start introducing radios for vehicle to vehicle (V2V) communications, there is great interest in determining how best to disseminate relevant content through such vehicular networks. Such vehicle-based dissemination offers the promise to offload significant portion of traffic from traditional cellular infrastructure. At the same time, it is of interest to see whether the pure vehicle-based dissemination could be improved by carefully leveraging the existing infrastructure in an efficient manner.

Due to the relatively small number of vehicles that may initially have such radios, as well as the high spatio-temporal variation of road traffic [3], vehicular networks are typically seen as an instance of intermittently connected mobile networks. A number of prior works have explored epidemic routing for content dissemination in intermittently connected mobile networks [1], [2]. However, as vehicular networks are likely to be operated in urban regions with significant cellular coverage, it is of interest to explore how the dissemination process could be improved by exploiting the cellular infrastructure to create an initial set of "seed" nodes.

The problem of seed-based dissemination of content in hybrid vehicular networks has been recently formulated and analyzed in [4]. The goal is to utility function that takes into account both the number of satisfied vehicles (i.e. those receiving the disseminated content by a given deadline) and the cost of seeding using the cellular network. In [4], the encounters for any two nodes was simplified to be identical and independently distributed. However, it is clear that the probability of two vehicles encountering each-other is related to their proximity, speed, and habits.

This work has been supported in part by NSF through award number 1217260

In this paper, seeding strategies that exploit these relationships are explored. First, a taxonomy of candidate strategies is presented. Then, example strategies and their algorithms are described. Finally, the strategies are simulated based on a set of real traces and their performance analyzed.

II. Taxonomy

One way to classify strategies is based on the feature they attempt to exploit:

Geographic Proximity strategies attempt to exploit the hypothesis that the probability of two vehicles encountering each-other within a given time budget is negatively correlated with their euclidean distance at the beginning of the dissemination time. *Speed* strategies attempt to exploit the hypothesis that the expected number of encounters for a given vehicle within a given time budget is positively correlated with that vehicle's speed at the beginning of the time budget. *Encounter Graph* strategies attempt to predict encounters within the time budget based on the frequency of encounters between vehicles. *Hybrid* strategies combine strategies from the other categories.

III. Problem Formulation

We assume that there is a particular configuration of vehicles at the initial time, and there is some given deadline or time budget. The relevant information about their positions, speed, etc. is assumed to be meta-data that is known to a centralized back-end server. The goal is to identify a set of vehicles which should act as seeds for epidemic propagation, i.e., to which the content to be disseminated should be downloaded through the cellular infrastructure. While the prior work [4] has investigated how to identify the initial number of seeds to optimize cost-delay tradeoffs, we assume in this work that this number has been fixed. We aim to maximize the number of vehicles that receive the content by the end of the given deadline.

IV. Geographic Proximity

A. Random Selection

Random Selection is simulated as the minimum acceptable performance and a reference for how well a strategy performs.

B. Most Neighbors

A node is likely to encounter its neighbors. This strategy chooses the nodes with the most neighbors.

C. Equal Sized Geographic Partitions

The area that contains satisfied nodes tends to expand out from the original seed's starting position. Distributing the seeds geographically distant from one another, the area of satisfied nodes from a given seed are less likely to overlap with another seed's area of satisfied nodes. The original region is divided into equal sized geographic regions, and the seeds are divided equally between these regions.

D. Neighbor Elimination

Choosing a seed that is nearby another seed is redundant, since it is likely to be seeded early on by the prior seed or to disseminate the file to the same group of nodes. A seed is selected at random without replacement from the set of candidate seeds, then the seed and all nodes within a specified distance (i.e. the seed's neighbors) are eliminated from consideration; this is repeated until all candidate seeds are eliminated.

E. Loner Elimination

Choosing a seed that is far from a other nodes is undesired, since it is unlikely to seed other nodes. All nodes with less than a specified number of neighbors within a specified distance are removed from consideration. The seed nodes are selected at random without replacement from the remaining nodes.

V. Speed

A. Speed Range Selection

Notionally, the faster a node is moving the more nodes it will encounter and the nodes encountered will be more geographically spread out than slower moving nodes. Speed is also likely correlated to the class of road a vehicle is driving on and how congested that road is at the time the speed is measured. Congested areas may be beneficial, because their are more nodes within proximity to disseminate the file. However, the lack of mobility may prohibit further dissemination. Nodes whose average speed is between a minimum speed and a maximum speed are selected as seed nodes.

VI. Encounter Graph

A. Encounter Probability Elimination

This strategy assumes vehicles that frequently encounter each-other in the past are more likely to encounter each-other going forward. A time-compressed weighted encounter graph is generated. Each vertex represents a node and the graph is fully connected. The weight of each edge is equal to the number of encounters between the two vertex nodes. The time of the encounter is irrelevant, hence the graph is time-compressed. A seed node is chosen from the remaining nodes at random without replacement. The nodes who are most likely to encounter the chosen seed are removed from consideration.

B. Highest Degree Selection

Highest Degree Selection selects the nodes who have the most incident edges in the *Encounter Graph*. By definition, these nodes encounter the highest number of distinct nodes. The dataset only contains one day of data, so the graph includes all timeslots in lieu of true historical data.

VII. Hybrid

A. Neighbor and Loner Elimination

For *Neighbor Elimination*, the pool becomes increasingly saturated with Loner nodes as the number of seeds is increased, because those near other nodes are eliminated. For *Loner Elimination*, the pool becomes increasingly saturated with neighbors of existing seeds, because those far from other nodes are eliminated. The two strategies complement each other and can thus be combined together. Seeds are selected at random if they have a minimum number of neighbors within a certain range, then their neighbors (potentially with respect to a different range) are eliminated.

B. Neighbor and Loner Elimination with Speed Range

This strategy tries to take advantage of both by selecting seeds with a minimum number of neighbors that are moving in a given speed range, then eliminating their neighbors.

C. Highest Degree Selection with Neighbor Elimination

At higher seed budgets, *Highest Degree Selection* performs worse than other strategies. This implies the highest degree nodes are well-connected, but isolated from other groups of nodes. By eliminating nearby nodes, this clustering effect might be mitigated.

VIII. Simulation Methodology

The strategies were evaluated using the same data set as in [4]. This data set contains GPS traces from taxis in Beijing. The data set spans twenty-four hours and each taxi reported its location in latitude and longitude coordinates at asynchronous one minute intervals. The simulation reused the normalized locations from [4], which approximates the location of taxis at synchronous one minute intervals via interpolation. The normalized data also excludes erroneous data points. The simulation limits itself to the same 632 well-connected nodes and encounter graph.

A *simulation* is provided with the strategy to use, the time of day to begin, and the number of seeds to use. After some initialization, the simulation invokes the strategy with the number of seeds. The strategy selects the nodes to use as seeds. Then, the simulation simulates the content dissemination for a period of 60 minutes at one-minute intervals. These intervals are referred to as *time-slots*. The number of satisfied nodes at each time-slot is recorded. Note that the simulation imposes two restrictions on the strategies. First, the strategy must use all the seeds. In some cases, strategies resort to *Random Selection* to expend unused seeds. Second, the strategy can only select seeds at the start of the simulation.

Simulations were run at start times ranging from 9 AM to 9 PM inclusive at one-hour intervals. The taxis were qualitatively observed via visualization to be more active during those hours. The strategies were tested with 3 different seed values: 10, 50, and 100. 10 and 100 were selected as extreme values and 50 as a data point in the middle.

For this paper, a *scenario* is defined as a tuple of data set, start time, number of seeds, strategy, and strategy parameters. Each scenario was simulated until the error of the sample mean for each time-slot was less than 1 satisfied node with 90% confidence. The sample means per time-slot were then averaged across the start-times.

IX. Evaluation

A. Geographic Proximity

1) Most Neighbors: Most Neighbors performs significantly worse than *Random Selection*. The neighbors of the node with the most neighbors often have the next highest number of neighbors, since they are in proximity to the neighbors of the node with the most neighbors. After 10 kilometers, the performance is inversely correlated to the range parameter, which increases the inclusion of nodes counting towards the number of neighbors.

2) Equal Sized Partitions: Equal Sized Partitions was simulated with two variants

- *Extrema*: extremes of latitude and longitude at the start of the simulation
- *Intelligent Borders(IB)*: a multiple of the standard deviation from mean of latitude and longitude at the start of the simulation

Intelligent Borders was simulated at 1 and 2 standard deviations. All variants were simulated with 10, 50, and 100 seeds. At all seed values, *Extrema* performed worse than random. At 10 seeds, *Intelligent Borders* with 2 standard deviations did marginally better than random, but at higher seed values it performed roughly the same as *Extrema*. *Intelligent Borders* with 1 standard deviation performed the best, but as the number of seeds increased its performance decreased. Around 100 seeds, its performance is marginally worse than random.

3) Neighbor Elimination: As its name implies, *Neighbor Elimination* is dependent on the definition of a neighbor. More precisely, the radius within which a node is considered a neighbor impacts the performance of the algorithm. To tune this parameter, the algorithm was simulated from 5000 meters to 65000 meters with 5000 meter steps and 100 seeds. A local maximum was observed at a range of 30 kilometers. Further refining of this parameter is unnecessary for the scope of this paper. *Neighbor Elimination* performs follows *Random Selection* at 10 seeds. The eliminated nodes make up a small portion of the overall pool to choose from, and so the strategy does not significantly affect the results. *Neighbor Elimination's* performance improves with the number of seeds. At 100 seeds, *Neighbor Elimination* satisfies on average 5 more nodes after 60 minutes.

4) Loner Elimination: To tune the range parameter of *Loner Elimination*, the algorithm was simulated from 5000 meters to 65000 meters with 5000 meter steps and 100 seeds. A local maximum was observed at a range of 30 kilometers. Further refinement of this parameter is not required for this paper. *Loner Elimination* outperforms *Random Selection* at all 3 simulated seed values (10, 50, 100). It performs best at 50 seeds. Notably, at 100 seeds *Loner Elimination* approaches the performance of *Random Selection* as the time slot increases.

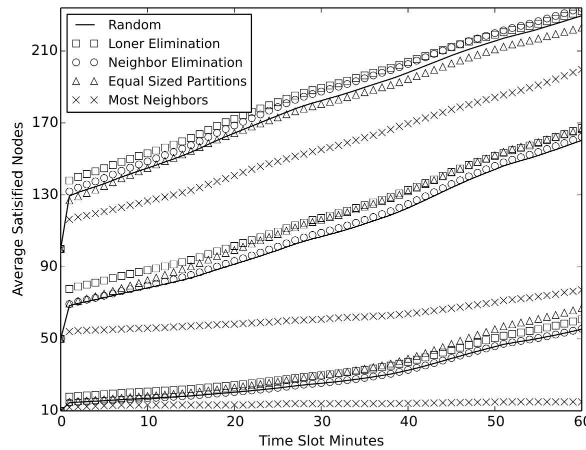

Fig. 1. Comparison of Geographic Strategies

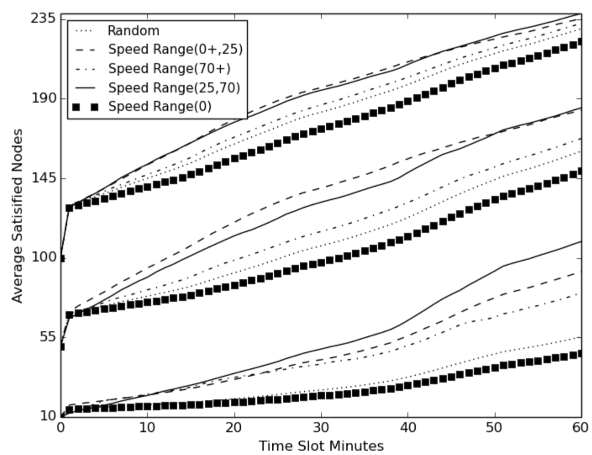

Fig. 2. Comparison of Speed Range Scenarios

B. Comparison of Geographic Strategies

Out of the Geographic Strategies, *Loner Elimination* performs the best with one exception. If given a low number of seeds and a long time budget, then *Equal Sized Partitions (Intelligent Borders, 1 stddev)* exceeds the performance of *Loner Elimination*.

C. Speed

1) Speed Range: Speed Range was simulated with 10, 50, and 100 seeds on the following speed ranges (in kmph):

Stopped	0 to 0
Slow	0 (exclusive) to 25
Medium	25 to 70
Fast	70 and up

Stopped performed worse than *Random Selection* at every seed value. This is expected, since stopped nodes are not exploring new geographic areas. Stopped nodes are also more likely to be in garages or otherwise isolated from normal traffic flows.

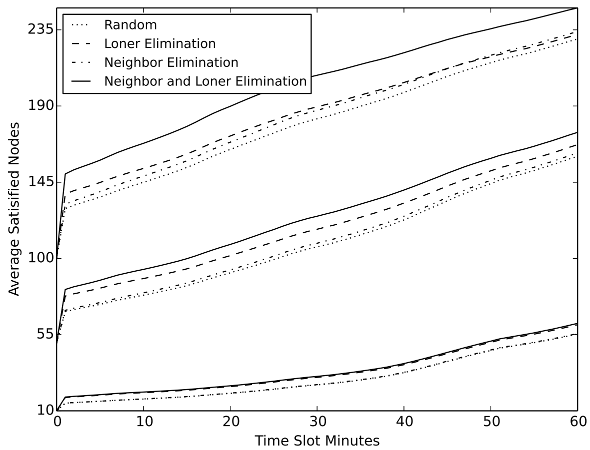

Fig. 3. Performance of Neighbor and Loner Elimination

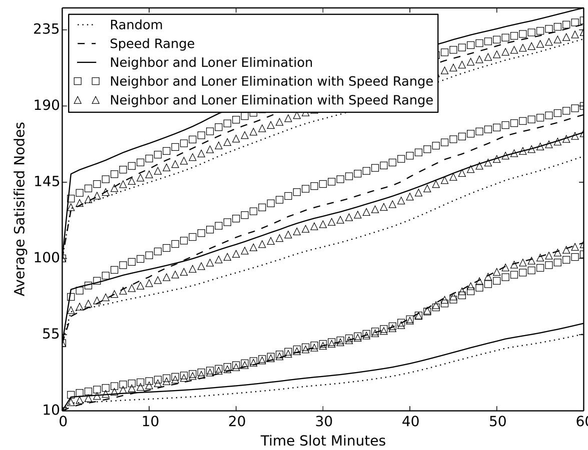

Fig. 4. Performance of Neighbor and Loner Elimination with Speed Range

The rest of the ranges performed better than *Random Selection*. *Medium* performed particularly well at 10 seeds, but performed roughly the same as *Slow* at 50 and 100 seeds. *Fast* was outperformed by *Slow* and *Medium*. *Fast* nodes may be isolated from other nodes, since vehicles moving at those speeds should be on freeways.

D. Encounter Graph

1) Encounter Probability: The performance of *Encounter Probability Elimination* was equal to *Random Selection* within the margin of error, and thus is not a strategy of interest.

E. Highest Degree Selection

Highest Degree Selection performs significantly better than *Random Selection* when allocated a small seed budget. However, its performance degrades as the seed budget increases. At 100 seeds, the performance was significantly worse than *Random Selection*.

F. Hybrid

1) Neighbor and Loner Elimination: Neighbor and Loner Elimination was simulated using the tuned range parameters of its namesake strategies (i.e *Neighbor Elimination* and *Loner Elimination*). At 10 Seeds, the strategy performs on par with *Loner Elimination*, which is better than *Random Selection* and *Neighbor Elimination*. At 50 Seeds, the strategy performs marginally better than all 3 strategies. It outperforms *Loner Elimination* by 7 satisfied nodes after 60 minutes. At 100 seeds, it performs significantly better than its namesake strategies. After 60 minutes, it has 12.5 more satisfied nodes than *Neighbor Elimination* and 18 more than *Random Selection*.

2) Neighbor and Loner Elimination with Speed Range: Neighbor and Loner Elimination with Speed Range Selection was simulated with the ranges tuned for *Neighbor Elimination* and *Loner Elimination*. Two variants were performed.

Optimal uses the optimal speed range from *Speed Range Selection*, which was *Speed Range Medium*.

Moving uses all speed ranges that performed better than *Random Selection*, which happens to be anything with a speed greater than 0.

This strategy performed better than Random at all 3 seed values (10, 50, and 100) for both variants. At 10 seeds, the strategy follows the performance *Speed Range(Medium)* closely. *Moving* performed slightly worse than *Optimal* at 10 seeds. At 50 Seeds, *Moving* performed significantly better than its component strategies except for the first few time slots, where *Neighbor and Loner Elimination* performs better. Limiting the speed ranges to *Medium* diminishes performance at 50 seeds, making this strategy worse than its component strategies. At 100 seeds, *Neighbor and Loner Elimination* performs better without including a speed range. Adding *Neighbor and Loner Elimination* to *Speed Range (Medium)* also diminishing performance, but when all speed ranges are included, it out performs *Speed Range(Medium)*. *Highest Degree with Neighbor Elimination* performed better than Highest Degree alone and better than Neighbor for lower seed budgets. Neighbor Elimination alone does better for long time budgets with high seed budgets.

G. Best Performers

Given a seed budget of 10 seeds, the best performer depends on the total time budget. *Highest Degree Selection with Neighbor Elimination* is best for time budgets up to approximately 15 seeds. *Neighbor and Loner Elimination with Speed Range* performs best of all strategies for time budgets greater than 15, but less than 40 minutes. *Speed Range (Medium)* performs best for longer time budgets.

Given a seed budget of 50 seeds, *Neighbor and Loner Elimination with Speed Range* performs the best of all strategies. *Speed Range (Slow)* performs the second best and nearly matches its performance for time budgets near 20 minutes.

Given a seed budget of 100 seeds, *Neighbor and Loner Elimination* (without Speed Range) performs the best at all time budgets.

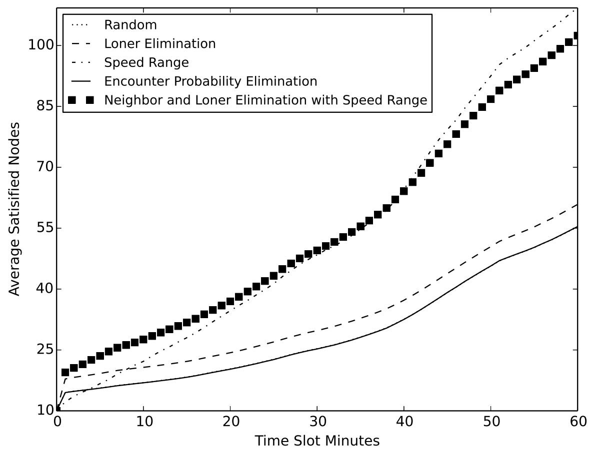

Fig. 5. Comparison of Best Strategies at 10 Seeds

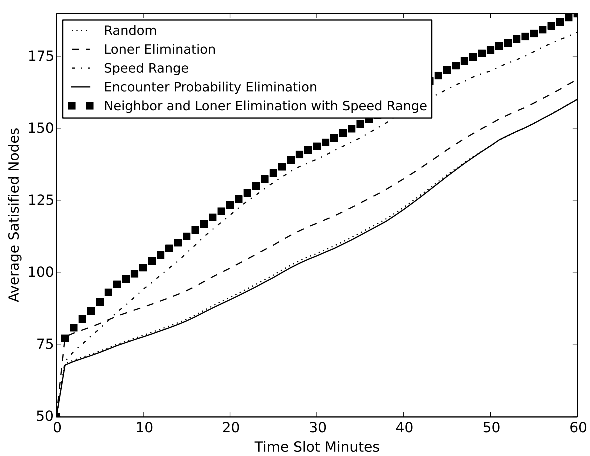

Fig. 6. Comparison of Best Strategies at 50 Seeds

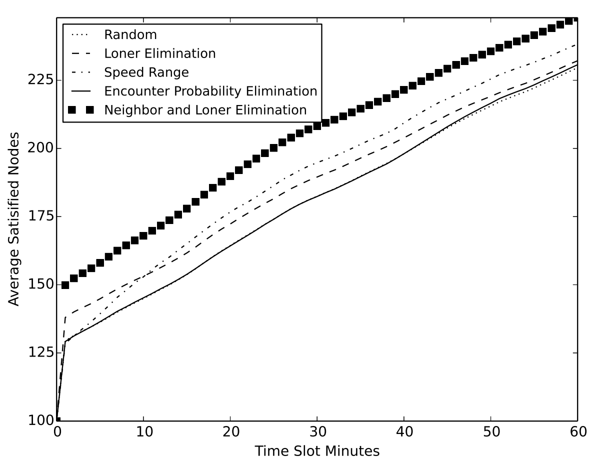

Fig. 7. Comparison of Best Strategies at 100 Seeds

X. CONCLUSION

Successful wide-spread dissemination of content in a Vehicular Network requires the content to be provided to a vehicle that comes in contact with many other vehicles. Several strategies for choosing seed nodes were defined, categorized, and simulated. We conclude from our experiments that speed-based strategies show the best performance when there is a low number of seeds and for longer time-budgets. When there are a relatively high concentration of seeds, then the best performance is obtained by employing speed-based strategy while simultaneously eliminating nodes that are too far from other nodes or too near other seeds.

REFERENCES

[1] A. Vahdat, D. Becker, "Epidemic routing for partially connected ad hoc networks", Technical Report CS-200006, 2000.

[2] A. Lindgren, A. Doria, and O. Schelen, "Probabilistic routing in intermittently connected networks", *SIGMOBILE Mobile Computing and Communication Review*, 2003.

[3] F. Bai and B. Krishnamachari, "Spatio-temporal variations of vehicle traffic in VANETs: facts and implications," *Vehicular Ad Hoc Networks*, 2009

[4] J. Ahn, M. Sathiamoorthy, B. Krishnamachari, F. Bai, L. Zhang, "Optimizing Content Dissemination in Vehicular Networks with Radio Heterogeneity," *IEEE Transactions on Mobile Computing*, 2014.