Ground-Truth Injection for Peer Grading using the Vancouver Algorithm

Donald "Drew" Bronson

2016-05-06

1 Simulation Parameters

The simulations below were run with the following default parameters:

- There are twenty group submissions.
- There are three students per group.
- Each student grades three assignments.
- The simulation is run two hundred times and the data aggregated.
- The legend indicates the number of ground-truth grades supplied to the algorithm.
- The default method for choosing which submissions ground truth grades are obtained for is to select them uniformly at random.
- Methods for choosing which submissions ground truth grades are obtained for are applied after ground truth is obtained for the entire planted cover. If the cover is smaller than the number of ground truth grades allowed, the grades used are chosen uniformly at random from those in the cover.
- Peer quality is represented by the number of draws a peer gets from a uniform distribution on the range (0, 1).
- Peer quality is uniformly random on the set 1, 2, 3, 4, 5.
- The true value of a submission's grade is always 0.5, the expectation of a uniform distribution on the range (0, 1).
- The grading algorithm used is the Vancouver algorithm, and it is terminated after ten iterations.
- The statistic plotted is the CDF of submission grade error, the quantity abs(submission grade from algorithm - 0.5).

- The default number of ground truths is the set 0, 5, 10, 15. All four of these CDFs are shown on each plot by default.
- Each plot runs its own batch of simulations.

2 Initial Verification of Vancouver Algorithm

The first thing to be done was to verify the number of iterations needed for Vancouver to converge to a reasonable output. Ten was hypothesized to be an appropriate number of iterations based on earlier simulations conducted by other members of the research group, and the simulations I conducted support this hypothesis. Figure 1 is the graph for ten iterations and Figure 2 is the graph for twenty iterations. As you can see, there is no significant difference between them. I have also included the graph for a single iteration, which represents simple averaging, in Figure 3. A comparison of the different numbers of steps is shown in Figure 4. This shows that for the default simulation parameters, simple averaging has comparable performance to Vancouver. Figures 1, 2, and 3 were run with the number of Vancouver iterations specified and the remaining settings default as described above. Figure 4 was run across several values for the number of Vancouver iterations, and five and fifteen ground truths. The similarity at all the various step values can be clearly seen by the degree to which the curves are similar. Figure 5 is included for the skeptic, and shows zero and ten ground truths over the same values for iterations.

3 Evaluation of Vancouver with Injected Ground Truths

Vancouver does not show better than linear decrease in error, which may be attributable partially or entirely to the linear decrease that is expected simply from adding in more ground truth grades. To understand this, consider that the error for a grade that is a ground-truth grade will always be zero. Therefore, as long as the grades which are not obtained from ground truth have some nonzero expected error k, adding one more ground truth grade will reduce the expected total error by k and therefore reduce the expected mean error by k/n, where n is the total number of submissions. Clearly, this is a linear decrease in mean error. Figure 6 is a plot of the expected mean error calculated from 100 trials with the other parameters default.

4 Evaluation of Vancouver with Grading a Cover

The simulations up until this point have assumed that a cover should be graded first before subsequent ground truths are chosen randomly to be fed into the algorithm, but this was not the initial case. Rather, in my first run of simulations, I chose assignments to inject ground truth for in a uniformly random manner,

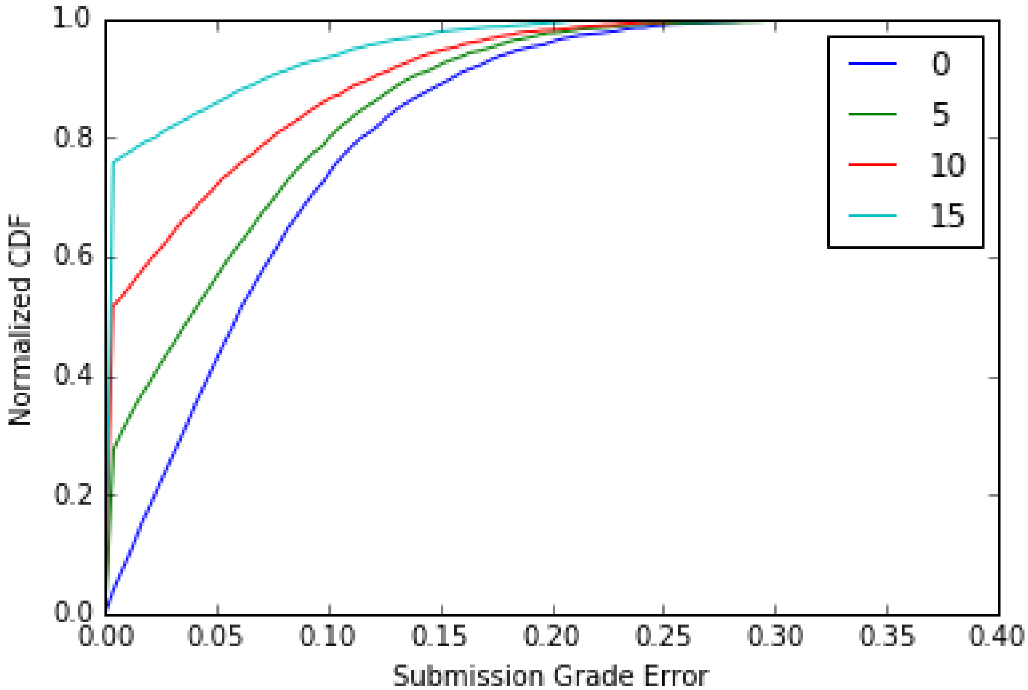


Figure 1: Ten-Iteration Vancouver

ignoring the existence of a cover. A plot of this is shown in Figure 7, which compares the two methods. This plot seems to indicate that there may be little to no benefit of grading a cover first, other than that it allows for grading of the graders. I also include plots of both of their CDFs for comparison to each other in Figures 8 and 9. Again, there is no discernible difference between the two figures.

5 Attempts to Find a Distribution for Which Ground Truth Injection is Better than Default

I next attempted to find a distribution of grader qualities for which ground truth injection would be more helpful than it is for the default uniform distribution of grader qualities on the set 1, 2, 3, 4, 5. Distributions that I tried included bimodal (Figure 10) and skewed both low and high (Figures 11 and 12). None of these distributions showed noticeable improvements over the default. It is worth noting for the sake of clarity that minor differences in the overall error are visible here, but this is due to the initial input from the graders being better or worse, since I was changing the distribution of quality of the graders. A comparison

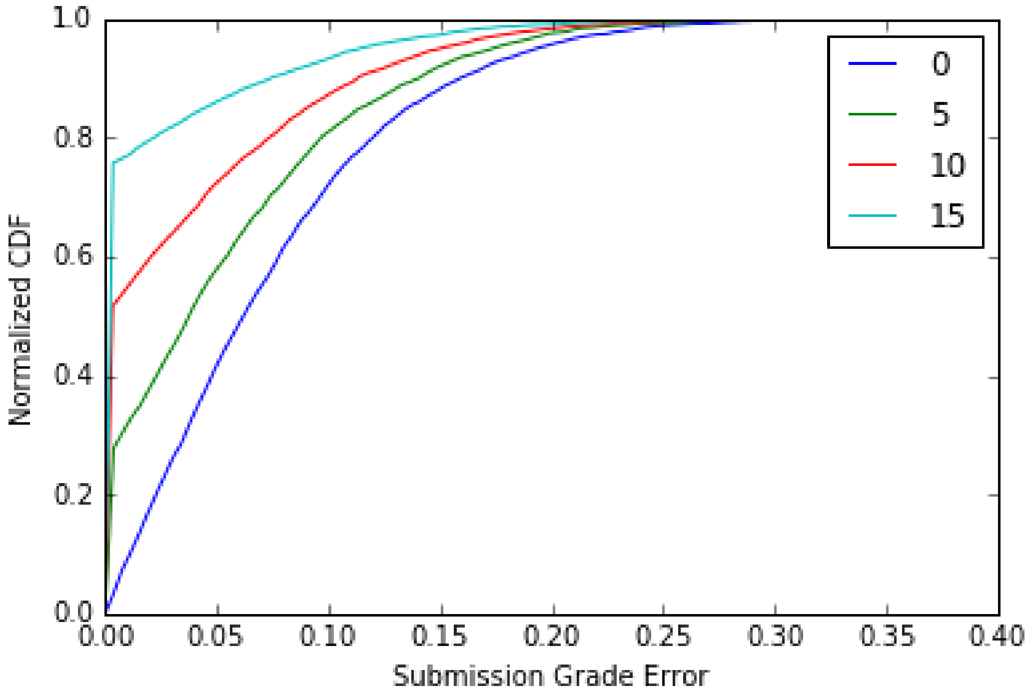


Figure 2: Twenty-Iteration Vancouver

of these at five, ten, and fifteen ground truths can be seen in Figures 13, 14, and 15, respectively. I have also, at the suggestion of a colleague for his best parameters for Vancouver, included the peer quality set {1, 1000}. From these plots it is clear that superior graders make for superior grades in the end. In the case where some graders are better than other graders by a phenomenal amount, Vancouver does exceptionally well.

6 Algorithms for Grading Past the Cover

Following the above examination, I looked into changing the algorithm for selecting submissions to provide ground truth grades for once the cover had already been graded. I tried two different types of algorithm: the first was greedy on the actual error in the grade (and thus omniscient), the second was greedy on the submission variance, as calculated by the algorithm. The latter should not have (and did not) show any improvement over random selection, as submission variance is not captured by our model, which treats all submissions as identical, and therefore the algorithm had nothing but noise to work with. The plot of this is shown in Figure 17. The first algorithm also did not noticeably improve the results (see Figure 16). A comparison of all three algorithms (including

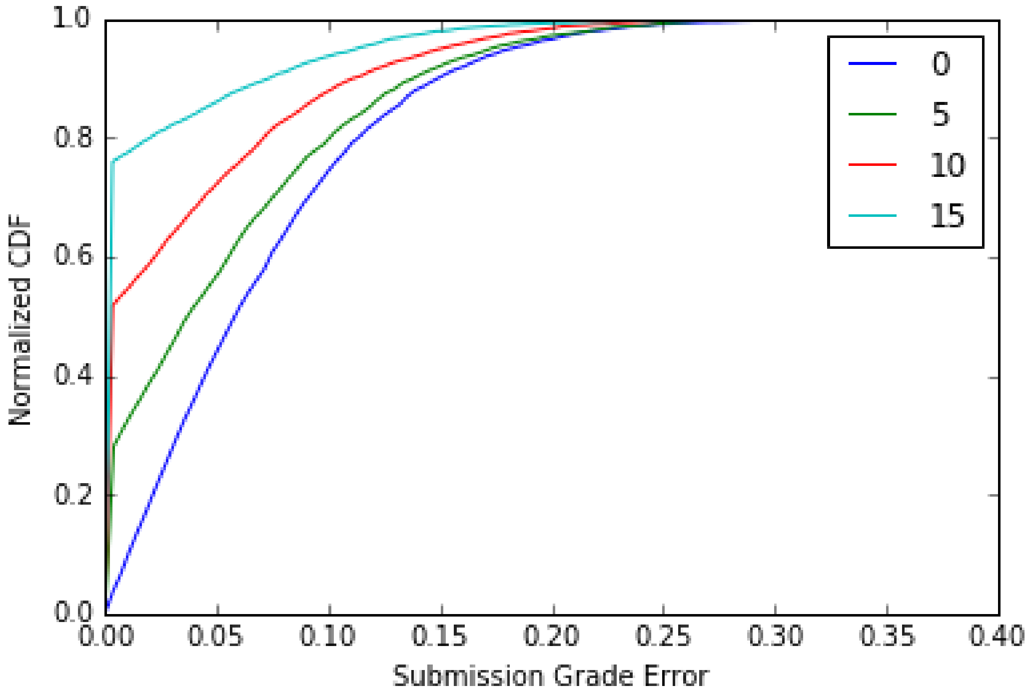


Figure 3: One-Iteration Vancouver (Simple Averaging)

the random baseline) is shown for various values of ground truth in Figures 18 and 19. This shows that for the default simulation parameters they are nearly identical.

7 Conclusion

Based on the results of these simulations, it does not appear that Vancouver's performance can be increased better than linearly by any of the methods for ground-truth-injection which were attempted. In fact, it appears that for these simulation parameters, Vancouver does not noticeably outperform simple averaging with ground-truth injection. Tangential results obtained by other members of the research group suggest that Vancouver does begin to outperform simple averaging at large enough sample sizes; however, these results were not run using comparable parameters and involved holding the number of students constant while the number of submissions increased to infinity, so they offer little bearing on this research. A possible future direction for research could be to re-run these simulations using larger sample sizes. A limited subset of these simulations was run with larger sample sizes and appeared to show no change, which is why such a direction was not pursued in this experiment.

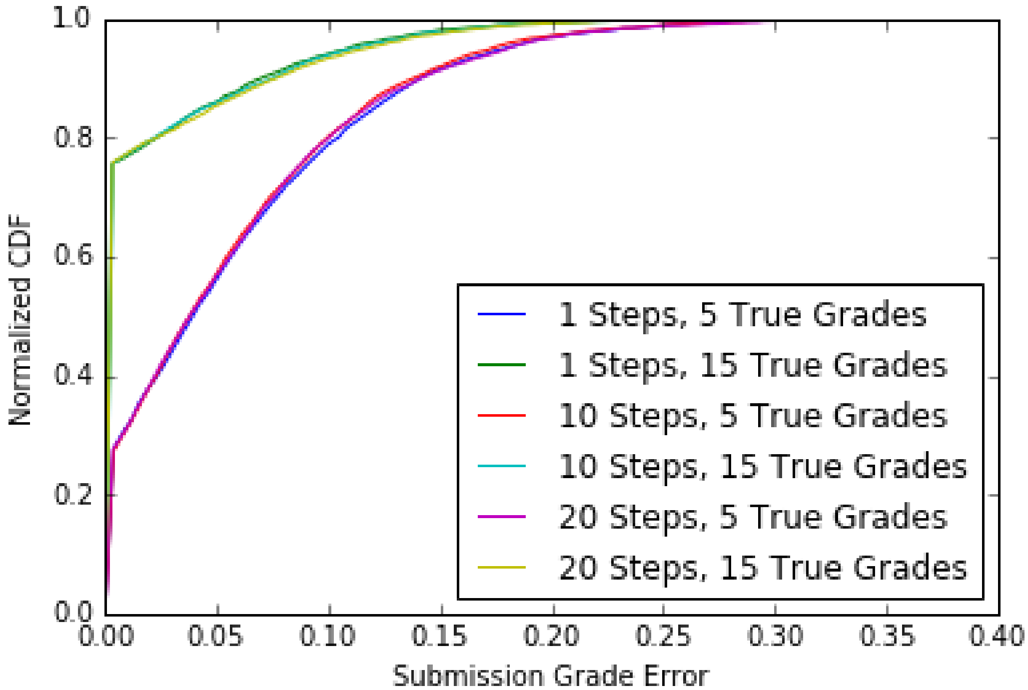


Figure 4: Vancouver vs. Iterations at Five and Fifteen Ground Truths

In addition, it is worth noting that the injection of ground truth may simply outweigh any differences in the algorithms used for obtaining the non-ground-truth grades, since ground truth so powerfully affects the average error.

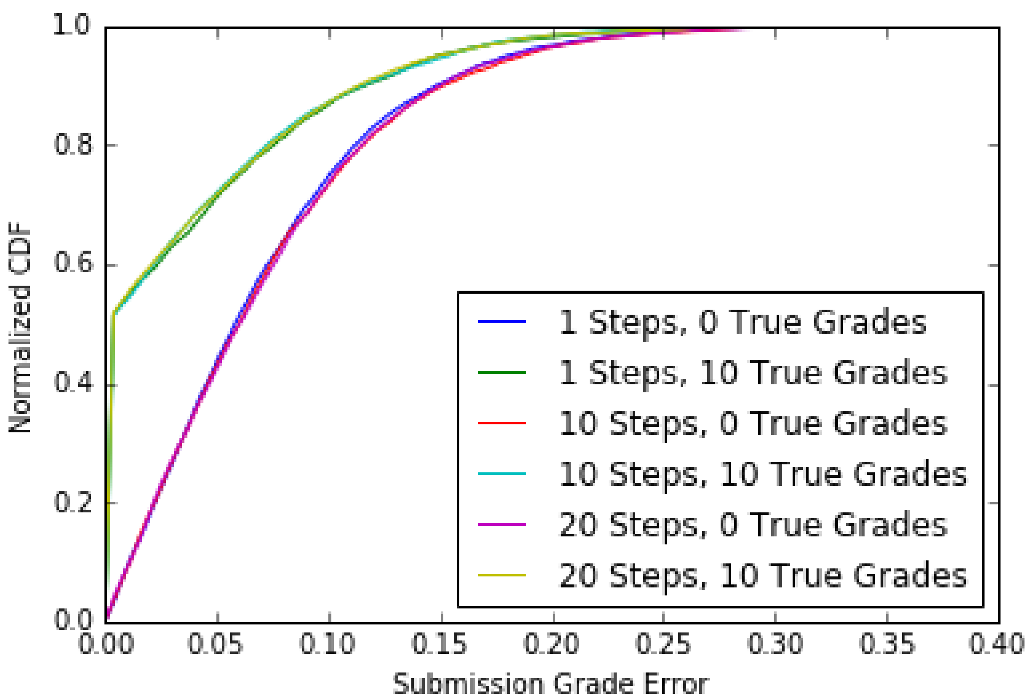


Figure 5: Vancouver vs. Iterations at Zero and Ten Ground Truths

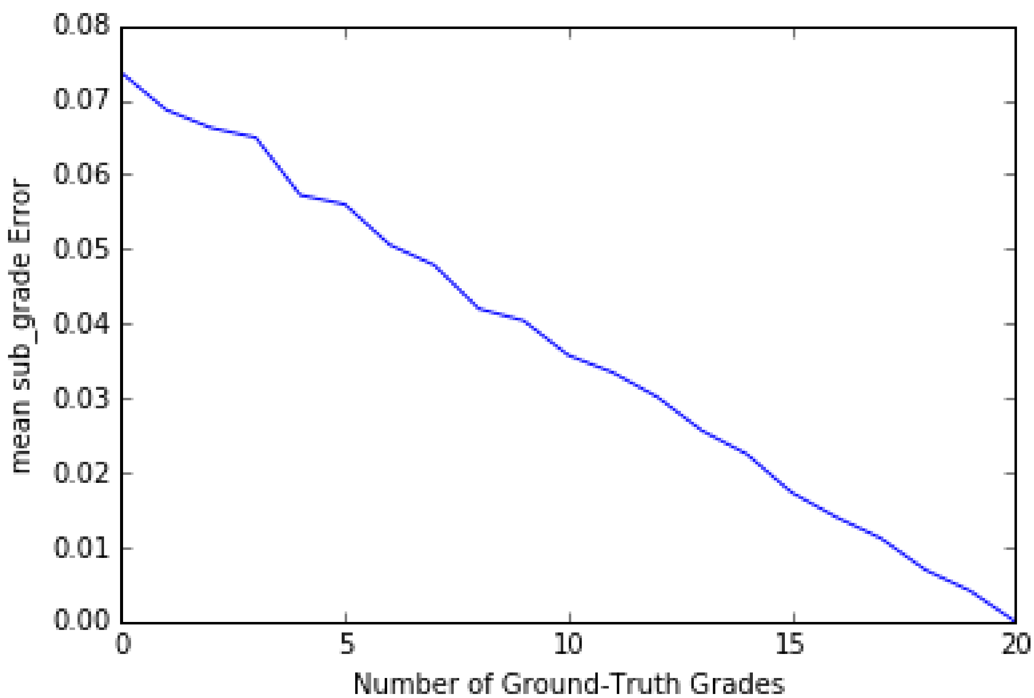
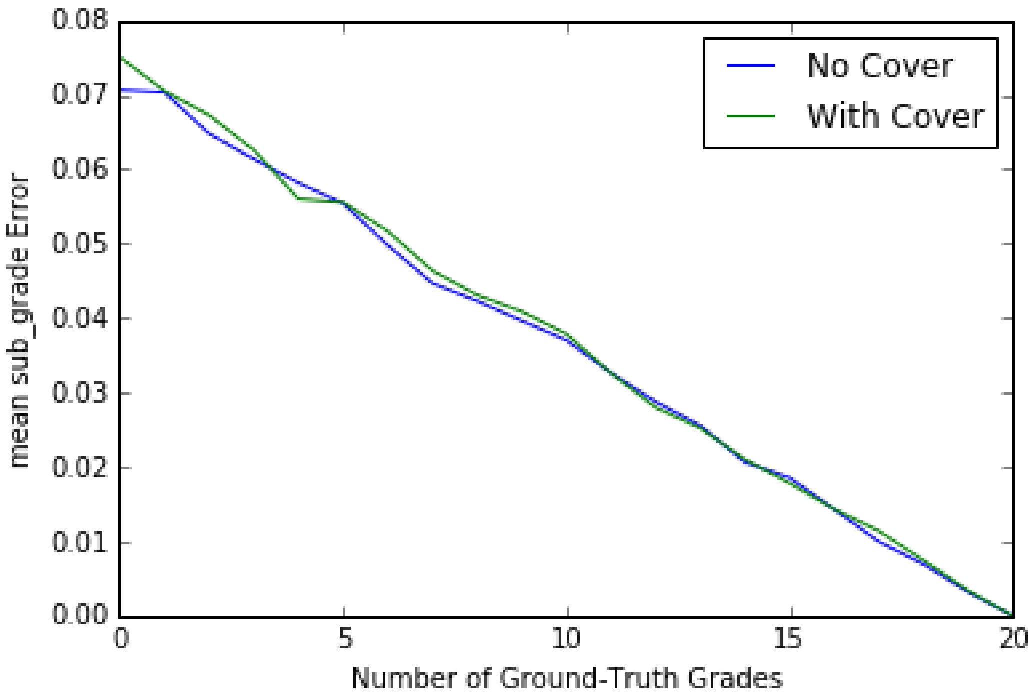


Figure 6: Vancouver as Number of Ground Truth Grades Increases



enddocument

Figure 7: Comparison of Vancouver With and Without Cover

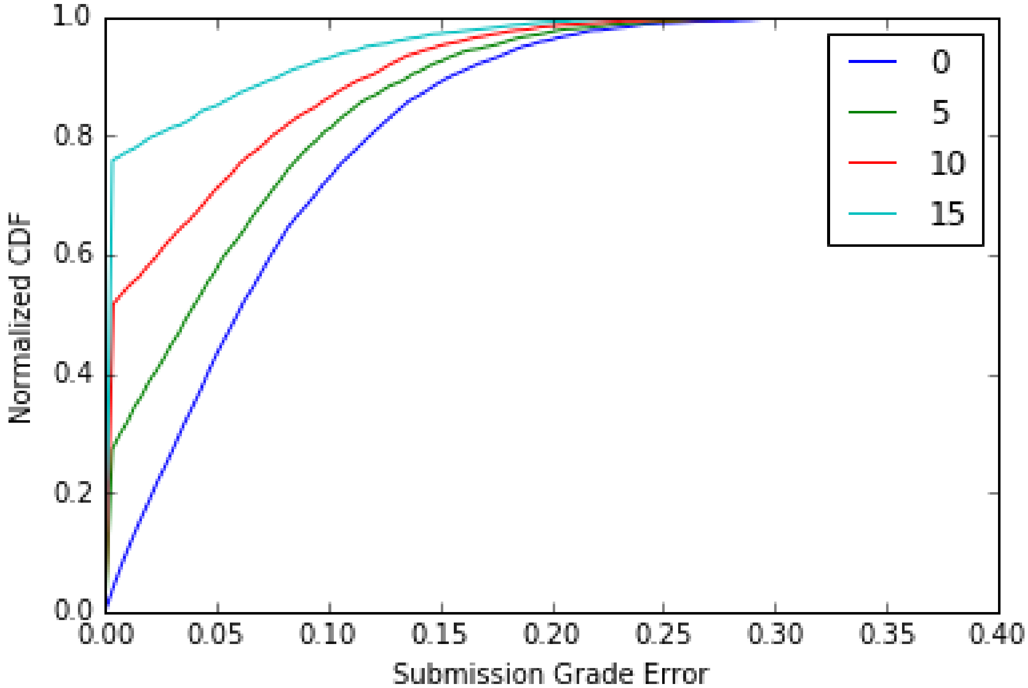


Figure 8: Vancouver (With Cover and Random Selection Past Cover)

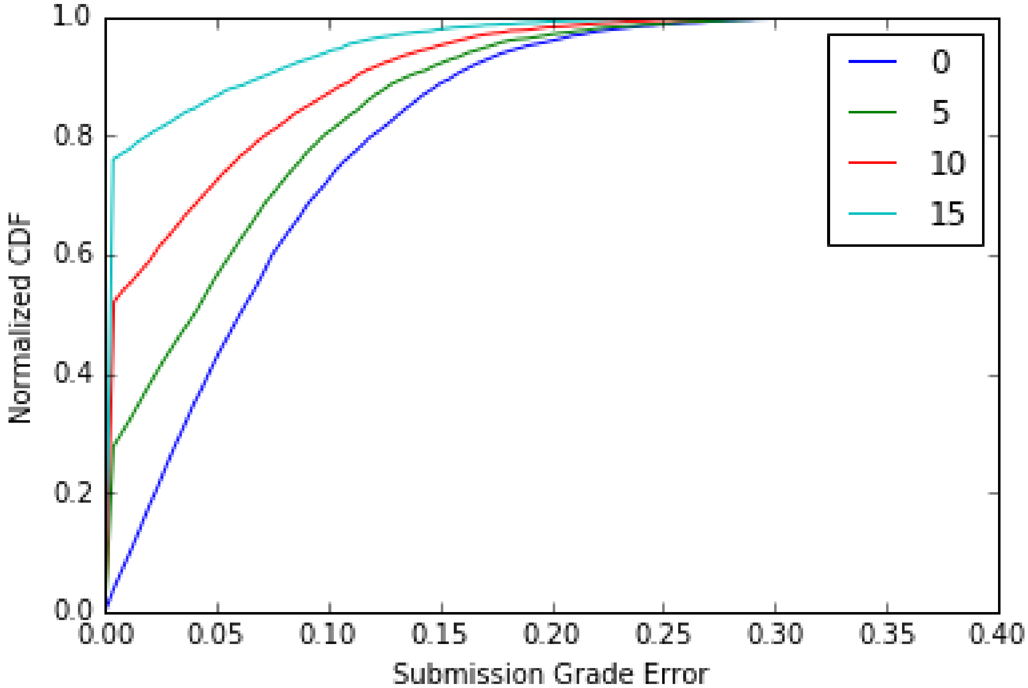


Figure 9: Vancouver (With Cover and Random Selection Past Cover)

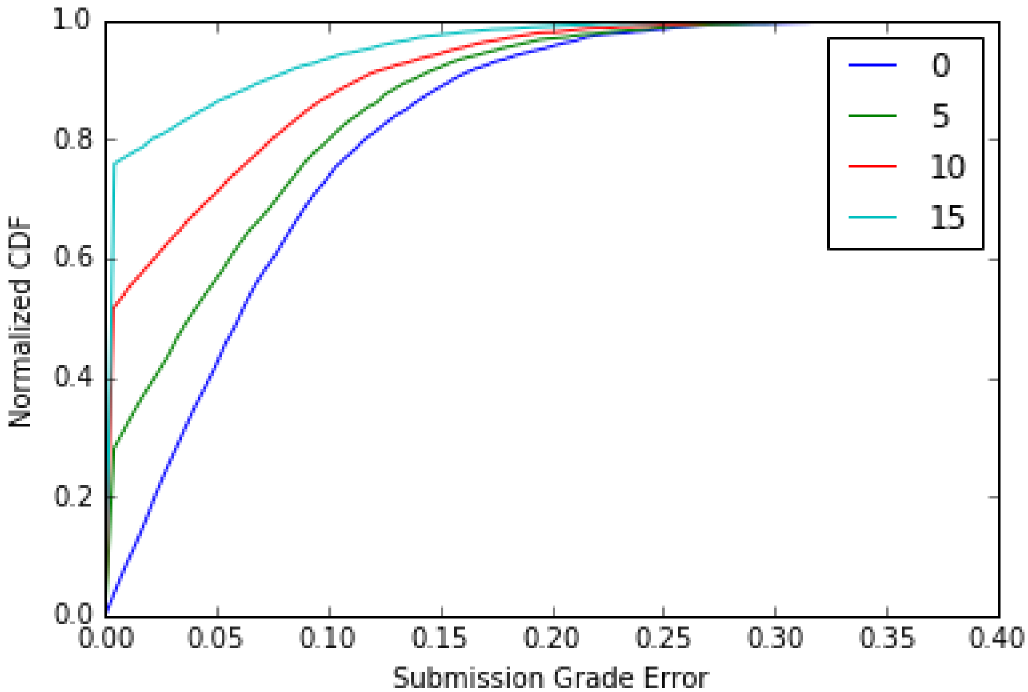


Figure 10: Vancouver (With Peer Quality Drawn Uniformly from the Set {1, 5}

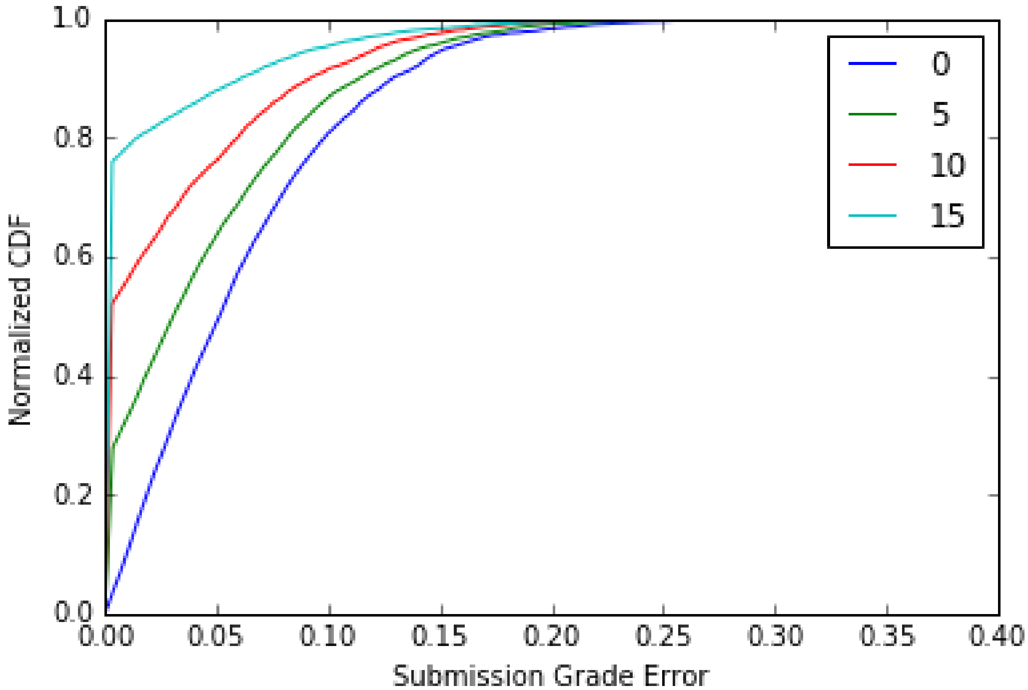


Figure 11: Vancouver (With Peer Quality Drawn Uniformly from the Set {1, 5, 5, 5})

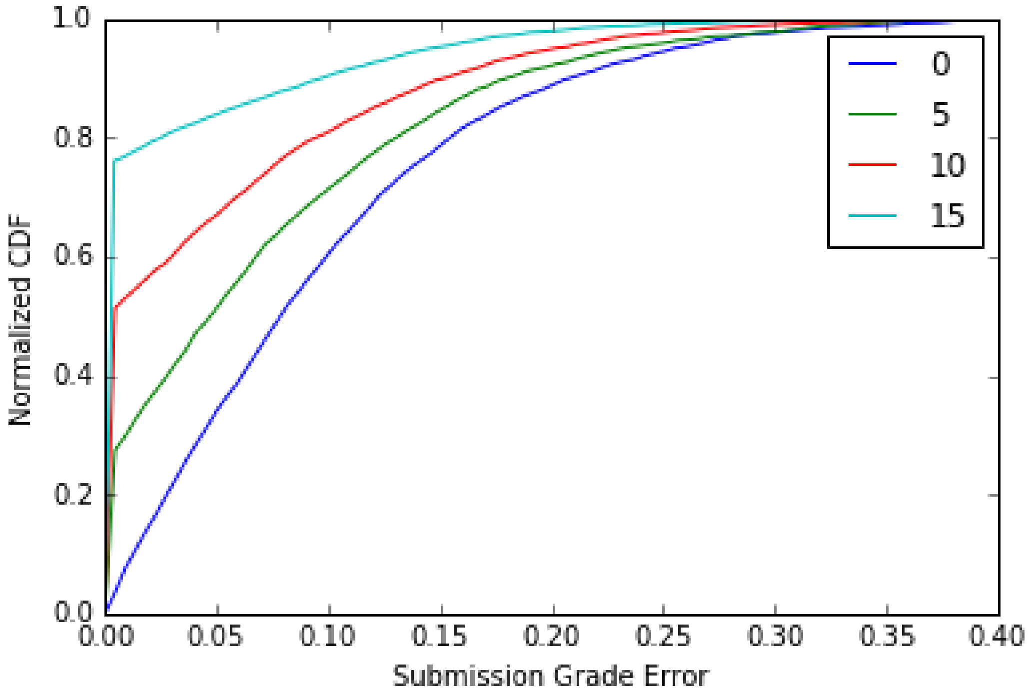


Figure 12: Vancouver (With Peer Quality Drawn Uniformly from the Set {1, 1, 1, 5})

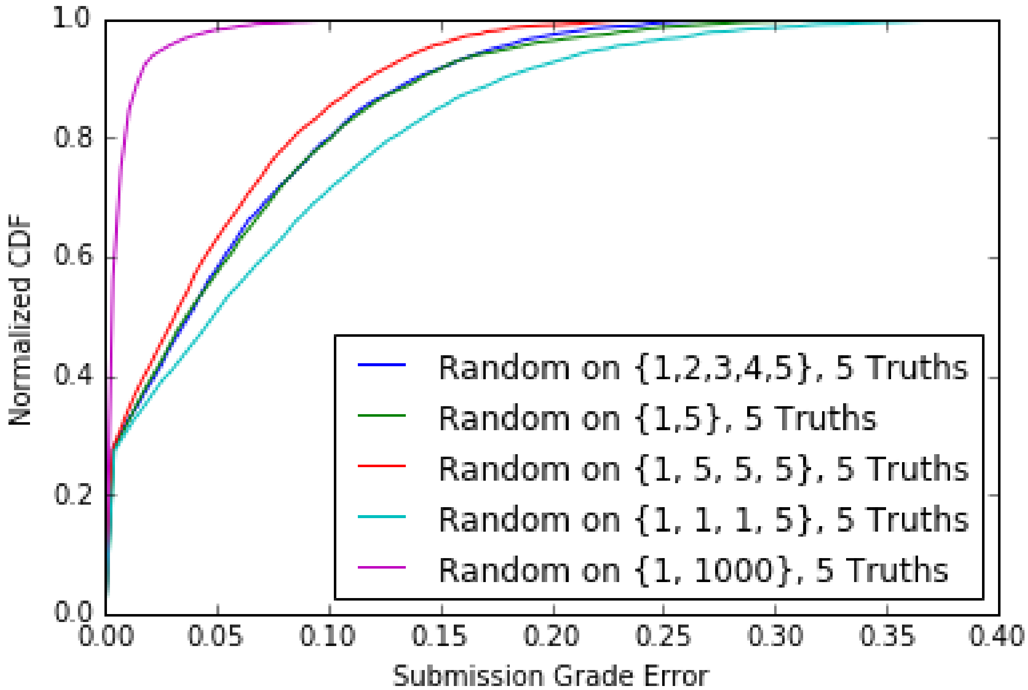


Figure 13: Vancouver with Varied Peer Quality Distributions, 5 Ground Truths

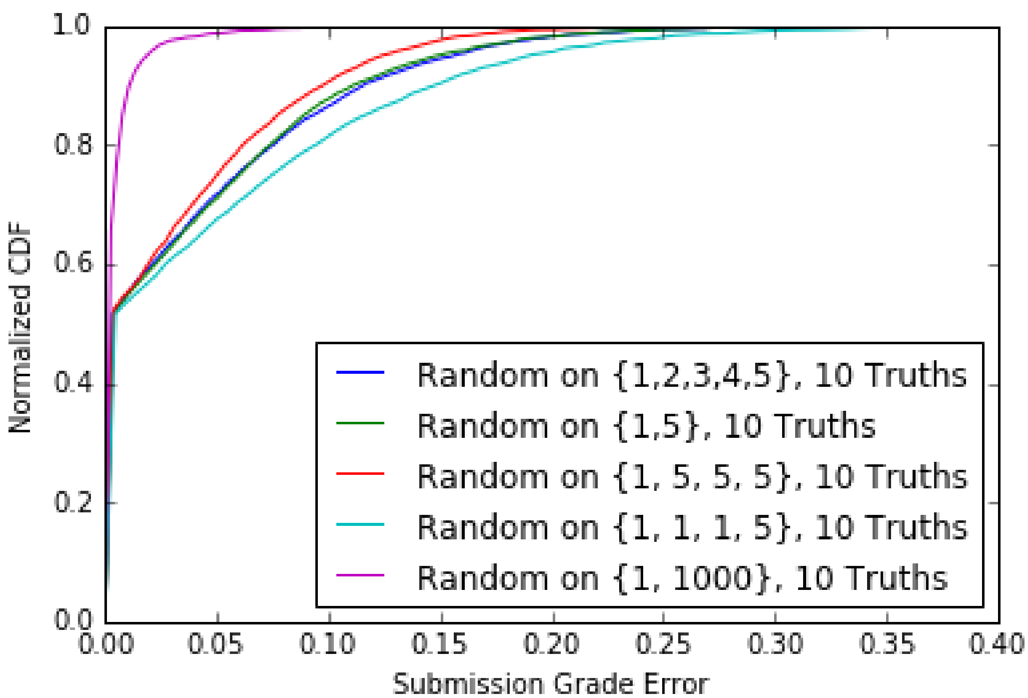


Figure 14: Vancouver with Varied Peer Quality Distributions, 10 Ground Truths

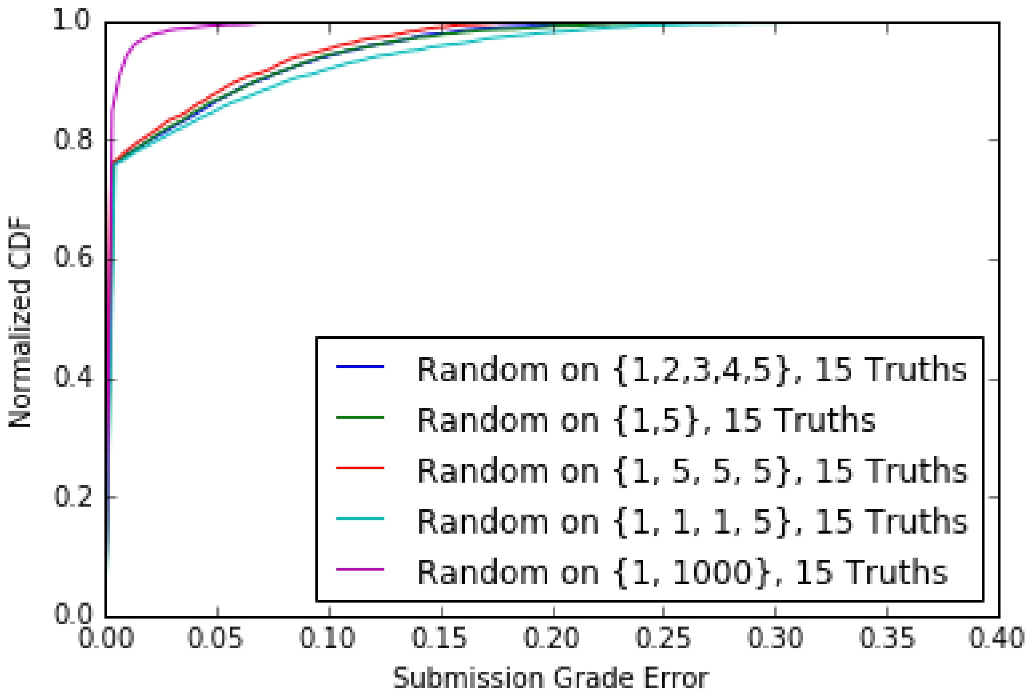


Figure 15: Vancouver with Varied Peer Quality Distributions, 15 Ground Truths

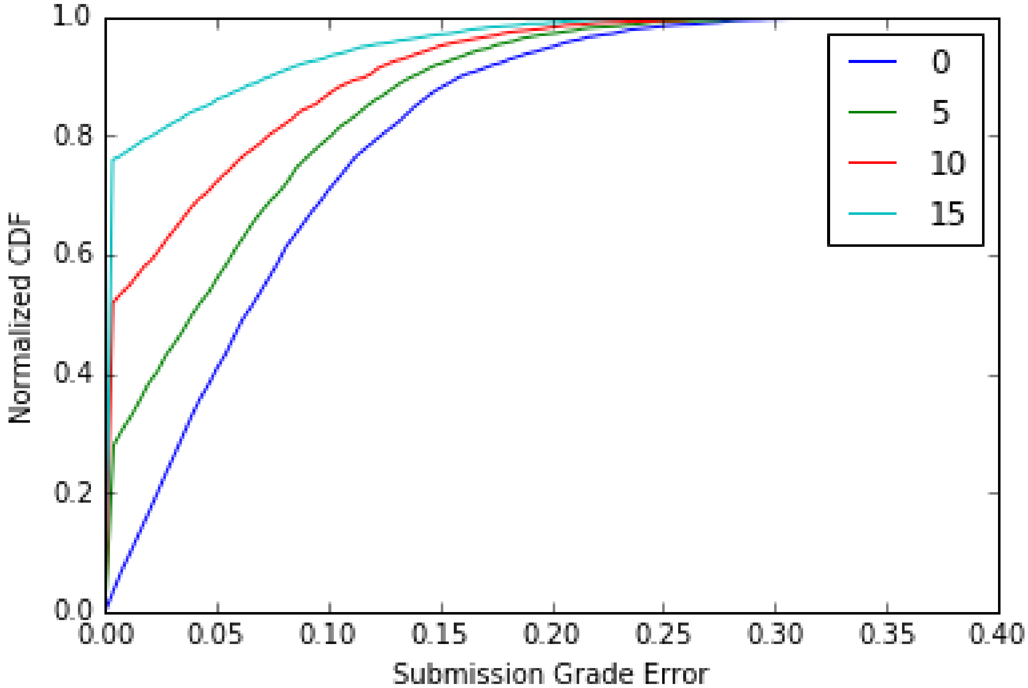


Figure 16: Vancouver with Greedy by Highest Grade Error (Omniscient)

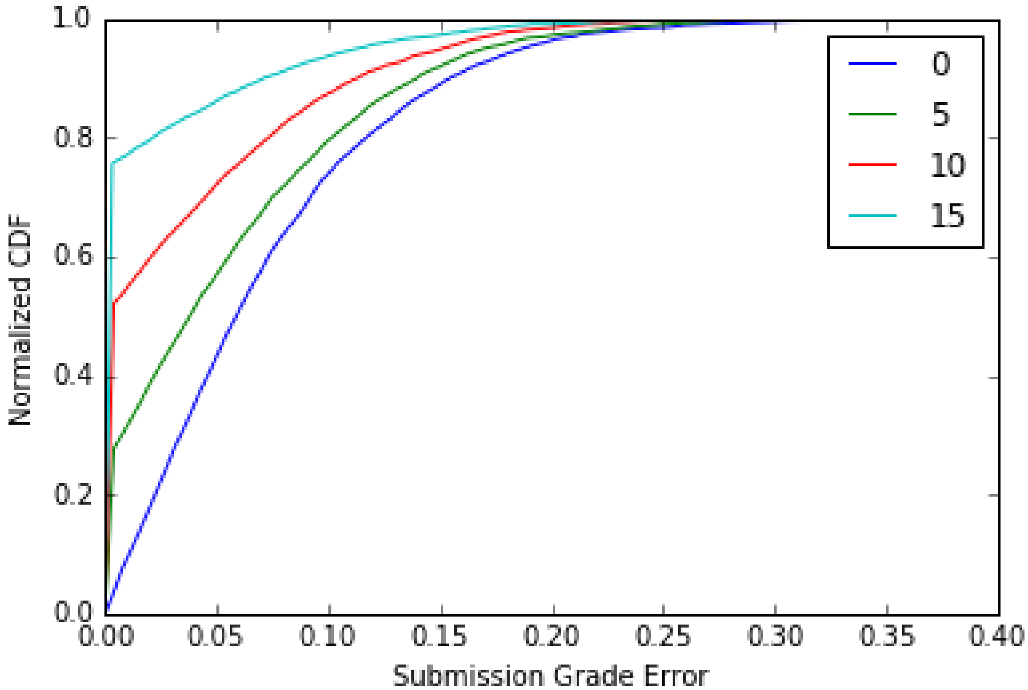


Figure 17: Vancouver with Greedy by Highest Submission Variance (Non-Omniscient)

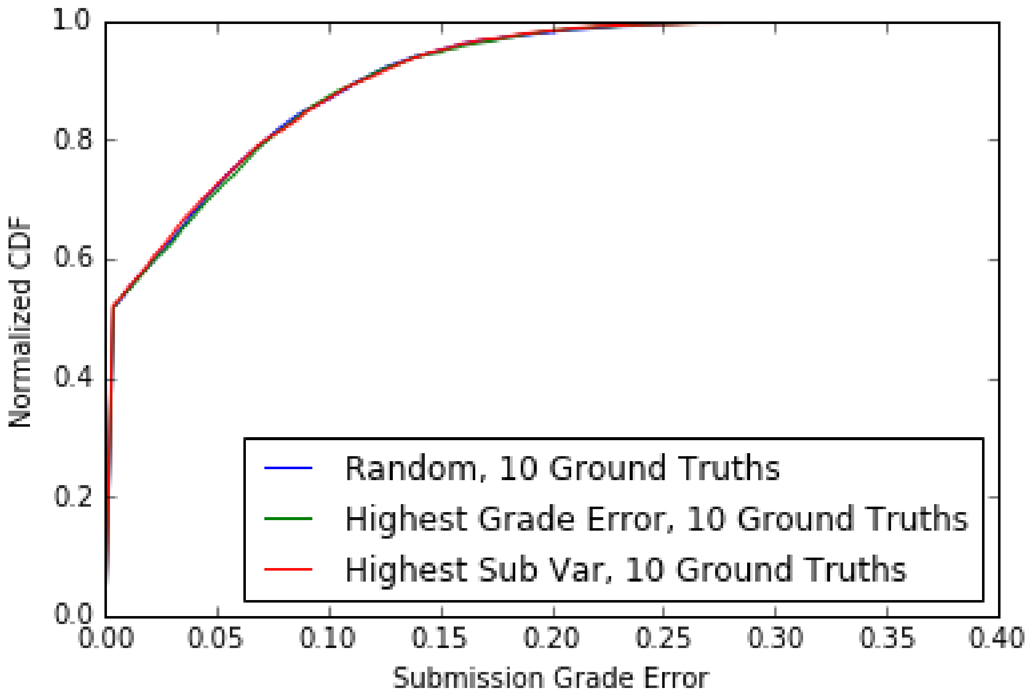


Figure 18: Vancouver with Various Algorithms, Ten Ground Truths

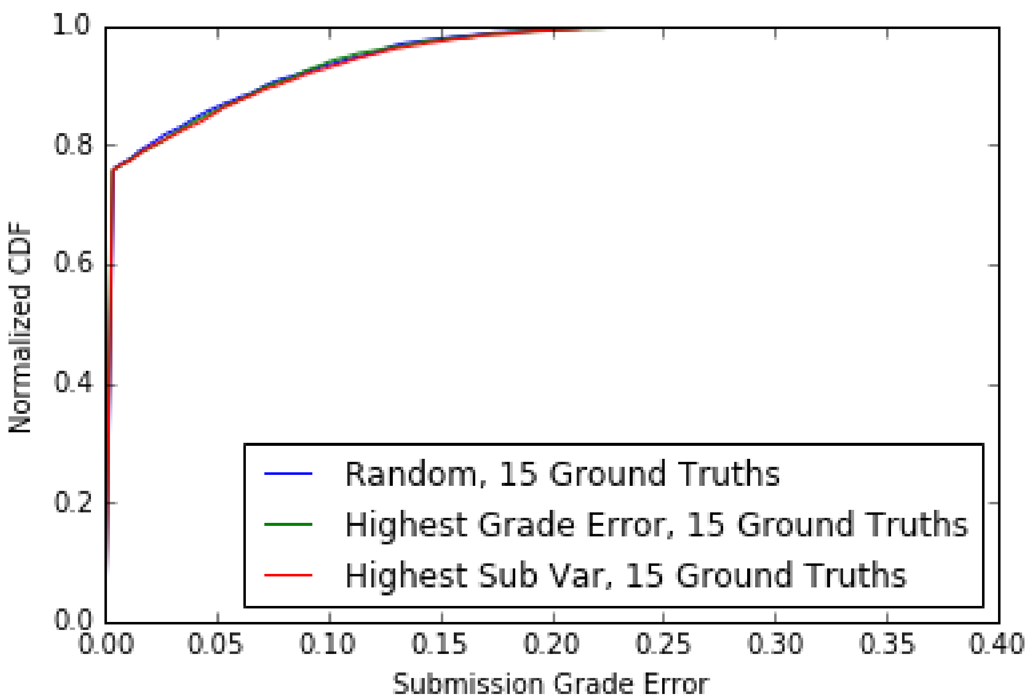


Figure 19: Vancouver with Various Algorithms, Fifteen Ground Truths