A simulation test bed for evaluating data analytic predictors of disinformation flow

No Author Given

No Institute Given

Data analytic methods hold great promise in utilizing real-world information to solve problems, but social systems offer substantial analytic challenges. These systems are often complex, have high variance, have limited potential for experimentation, and often lack accessible ground truth. Recent work has shown that social simulations can help to address these challenges by serving as test beds for data analytic methods. These test beds generate synthetic data based on known and adjustable underlying causal relationships, allowing evaluation of data analytic methods under dataset shift. Synthetic data from test beds can also supplement real-world data to incrementally increase the the complexity of training data, potentially enhancing performance and robustness.

We have developed and implemented a simulation test bed that generates synthetic data for assessing data analytic methods used to predict the flow of disinformation. The test bed uses an agent-based information diffusion model that integrates multiple diffusion cascades, agent attention limitations, and a complex social network. We generated 45 synthetic data sets from the test bed, $S_0 \ldots S_{45}$, with varying parameter settings that generate significantly different numbers and lengths of information cascades.

We used the simulation test bed to evaluate the predictive performance of two data analytic methods: ridge regression and decision trees (chosen for their different learning capacities). For every pair of synthetic datasets (S_i, S_j), the two machine learning algorithms were trained (on S_i) and tested (on S_j) to predict the eventual size of a disinformation cascade based on data from the first 5 timesteps. By training and testing on different synthetic data sets (drawn from different simulation parameter settings), we can evaluate the ability of the two algorithms to address dataset shift.

We define performance based on the mean relative performance of the algorithm over all pairs of synthetic data sets: $\omega = \frac{\text{mean(performance(Train on } S_i\text{, test on } S_j\text{))}}{\text{mean(performance(Train on } S_i\text{, test on } S_j\text{))}}, \forall_{i,j}$. The closer this metric is to 1.0, the better the method performs in the face of dataset shift. Table 1 shows ω for two evaluated algorithms.

As expected, decision trees (the higher capacity algorithm) perform nearly 4 times better than ridge regression. Further results (omitted due to space) indicate that performance is non-uniform across training/testing data sets. Training on certain data sets provides better average performance, which suggests that using synthetic training data of different complexities would enhance robustness.

Table 1. Performance of data analytic methods under dataset shift

	ω
Ridge Regression	5.44
Decision Tree	1.38