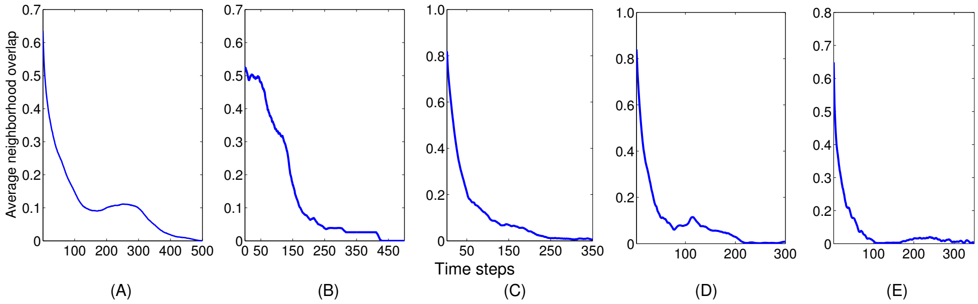

Fig. 6. The average neighborhood-overlap value at different lags for (A)INFOCOM 2006, (B)SIGCOMM 2009, (C)Highschool 2012, (D)Highschool 2011 and (E)Hospital datasets.

each active node at time t the overlap in its neighborhood between two time points. To measure this overlap we use the standard Jaccard similarity as has been pointed out in [2]. Note that this is one of the most standard and interpretable ways to measure structural similarity as has been identified in the literature with applications ranging from measuring keyword similarity [29] to similarity search in locality-sensitive-hashing (LSH) [30]. It has also been extensively used in link prediction [31,32] as well as community detection [33]. Figure 5 shows how we formulate this measure using the Jaccard similarity index. We represent the neighborhood overlap at lag k as the mean value across all the active nodes in time step t. To measure the extent of similarity we measure neighborhood-overlap for each snapshot at different lags and take the average of them. This essentially shows, given a time specific snapshot how the similarity changes as we increase the lag. Figures 6(A) - (E) show how this similarity changes with time as we increase lag for the five different datasets. As we increase the lag the similarity decreases almost exponentially and hence considering snapshots at larger lag where the similarity value is very low could introduce error in learning the auto-regressive equation. Also for a higher similarity value the corresponding lag would increasingly introduce more error in the fit due to lesser number of data points on which the ARIMA model gets trained to learn the fit function (see figure 7). In fact we observed that the error in prediction increases if we consider a lag too small (high similarity value) or too large (low similarity value) (see figure 7). Hence we consider the similarity value of 0.2 as the threshold for calculating the lag. For our prediction framework the corresponding value of the lag acts as the window for fitting the ARIMA model.

Let the size of the window be w and we want to predict the value of the time series at time t. To our aim we consider the time series of the previous w time steps consisting of the values between time steps $t-1-w$ to $t-1$ and fit the ARIMA model to it and obtain its value at time step t. We repeat this procedure for forecasting at every value of t. Thus, the time step t is the test point and the series of points $t-w-1$ to $t-1$ form the training set. One can imagine this process as a sliding window of size w which is used for learning the auto-regressive equation and the point that falls immediately outside the window is the unknown that is to be predicted.

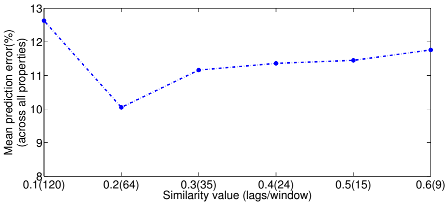

Fig. 7. Mean prediction error (%) across different properties for INFOCOM 2006 dataset for different similarity values. The lags corresponding to the similarity value are also provided.

7 Prediction results

In this section, we provide detailed results of the our prediction framework on the datasets discussed earlier. To determine the accuracy of our prediction strategy we use the cross validation technique. For each time step in this range we use our framework to obtain a prediction at that time step. Since we already know the original value, we can obtain a percentage error for the prediction. Let $predict_t$ represent the prediction value at time t and $original_t$ represent the original value. We obtain percentage error ($error_t$) using the formula:

$$error_t = \frac{|original_t - predict_t|}{original_t} * 100$$

First we try to find the suitable window for predicting the value of a time series at a time step. For this we refer to figure 6 where we quantify structural correlation and show how the similarity value decreases with increasing lag. We observe that the value of the structural correlation decreases as we increase the lag. For INFOCOM 2006 dataset (figure 6(A)) the correlation drops to less than 0.2 at lag around 70. Therefore we select a window of size 64. We could have selected any other value between 60 and 70, but we select 64 as it is in the power of 2 and it helps in the spectrogram analysis. Similarly we find the