# Finding an Optimal Subset of Training Parameters for Studying Civic Engagement with Link Prediction

LAUREN LAM, ALEX LOEWI
GROUP 14

## ABSTRACT

A data set collected by Stanford Professor of Sociology Dan McFarland, containing records of interactions between high school students and a large number of attributes pertaining to each student individually was analyzed with NetworkX and the Python packages Scipy and Numpy. Link prediction was used to explore how past interactions and these attributes correlated with future interactions Given the large number of attributes and the effort involved in both collecting and analyzing them, a subset optimized for size and predictive accuracy was searched for. This was done by creating power sets of the attributes and using the Murata weighted link equation multiplied by the Adamic/Adar score to predict future interactions. In general, it was found that the attribute-groups (such as gender. membership on the the football team or in the French club) that most increased the predictive accuracy from a baseline purely of past interactions were those that had the greatest number of members (see Appendix). These results were then normalized based on group size, but the method give results that were difficult to interpret. The optimal number of attributes (including both group membership and demographic information) for prediction was found to be four.

## 1. INTRODUCTION

The work is inspired by a number of sociological studies, the most influential of which center around "social capital," defined roughly as the value of the social relationships within a population. Social capital is closely linked with a large number of important measures, such as health, economic success, educational success, and low crime rates. One of the correlations that has gotten the most academic attention is social capital's impact on civic engagement: the amount the people take part in their community, help their neighbors, volunteer their time for charitable causes, and participate in politics at both the local and national levels.

It has been shown that social capital is well predicted from rates of group membership in a community, whether those groups are Parent Teacher Associations, workers' unions, or recreational sports leagues. However the original work on the topic did not delve in detail into the different effects of the different types of groups. Further work was able to examine the effects of particular types of groups by looking at high school student groups, such as quiz bowl and French club. The results indicated a strongest positive correlation with service and politics oriented clubs, but also with arts groups such as music and drama. Academic clubs and sports had a range of effects, both positive and negative.

Given the importance of groups to social capital, but also the large number of parameters involved in the study done on student groups, it seemed that it would be helpful to future work to find the smallest possible set of parameters---such as the number of interactions between students, ethnicity, household income, or membership in other groups---that would accurately predict future interactions or group membership. This dataset included how many interactions had been witnessed occurring between pairs of students over the course of a year, with which groups these students were involved, and demographic information about the students.

## 2. ANALYSIS

### 2.1 The Data in the Graphs

The graphs on which the present state of the link-prediction algorithm have been run have been built from the interaction records, and the group membership records. Each individual is represented by a node, and each interaction is represented by a separate time-stamped link between two people. The time stamp is used to choose a chronologically meaningful training set for the link prediction. The data set also contains a detailed profile on each student including their membership status in various organizations and groups.

### 2.2 Network Structure

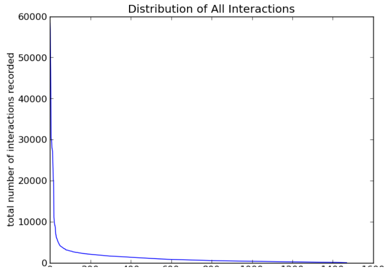
#### 2.2.1 Clusters

The graph was formed first only from interactions, and then all students in a given group were made into a clique (without looking at the edges already there) with edges tagged with the name of the group. Clustering coefficient and average path length algorithms were then run on both of these graphs to see if there was any difference---there was not. Knowing that the clustering and path lengths do not change based on the groups, we could then think about the group information as simply augmentation of the results of the interactions. If they were disjoint sets of edges, then it might have proved difficult to use both in a meaningful way as predictive attributes.

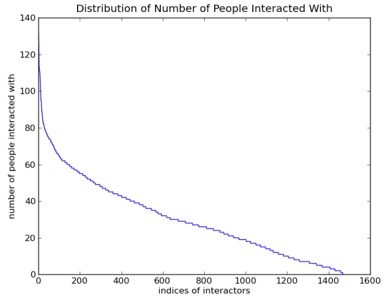
| | Clustering Coefficient | Average Path Length |
|---|---|---|
| Interactions Only | 0.57709 | 3.859768 |
| Interactions and Groups | 0.57709 | 3.859768 |

### 2.2.2 Distributions

The data was used to form several slightly different networks, which either revealed slightly different things about the data, or were the only tractable form in which a particular trait could be analyzed.



The distribution of the total number of *interactions* per individual in the data set was the familiar power-law-like curve---intuitively predictable from the "rich get richer" model of network growth. Social skills and a well-maintained social position require a lot of work, but pay large dividends and self-reinforce.



However the distribution of the total number of *distinct individuals* talked to over the course of the year had a distinctly different shape, seemingly much closer to linear.. One hypothesis regarding the reasons for this was that it was a function of the sample size, and length of the period observed. Given that the schools were an essentially closed environment and that the pool of people with whom to interact was also relatively small (especially when considered in the context of having a full year to interact with them ) it seems initially possible that the distribution is simply not getting enough "space" to form---if the populations were unbounded, then perhaps the more socially precocious individuals would out-perform the rest of the population in ways in which they cannot here. It is worth noting that the maximum degree node on the graph of unique individuals is only around 120, far below the total number of people in the graph---but this may only be reflective of the fact that the population is also strongly bounded in smaller subsets, such as by grade, or class---a phenomenon well in line with anecdotal experience from high school. It therefore does not allow for the dismissal of the hypothesis that this type of graph might be a particular signature of tightly circumscribed groups.

### 2.3 Link Prediction Algorithm

The attribute based link prediction algorithm works by computing a score based on the training range data for each pair of nodes in the network and then using these scores to predict the links in the testing range of the data.

First, we implemented the Murata link prediction algorithm and run it using various attributes in our data. The algorithm takes every pair of nodes in the graph and determines a score representing the likelihood of an edge between them being formed. This score is computed using the Murata weighted link equation multiplied by the Adamic/Adar score. The Adamic/Adar score is the inverse of the sum of the degree of the common neighbors, which is designed to give more weight to those neighbors which are less common.

Adamic Adar:

$$score(x,y) = \sum_{z \in N(x) \cap z \in N(y)} \frac{1}{log\,|N(z)}$$ ,

where N(x) denotes the neighbors of x

Murata:

$$score(x,y) = \sum_{z \in N(x) \cap z \in N(y)} \frac{1}{log\,|N(z)|} \times \frac{weight(x,y) + weight(y,z)}{2}$$

,

where weight(x,y) is the # of interactions of x and y

Then, we extended the Murata formula to factor in the number of common groups. We chose to multiply this factor into the score so that the factor has the same relative impact on all the weights.
The foruma multiplies the number of times the the two individuals interacted, multiplied by the number of groups within the set s that both x and y are members of.

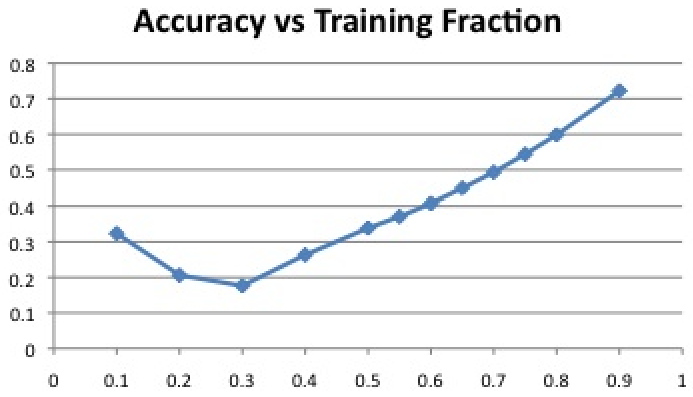
Attribute Set Weight:

$$weight(x,y,s) = TimesInteracted(x,y) * \sum_{x,y \in g,\ g \in s} 1$$

where x and y are two distinct people, s is a set of attributes

2.4 Experiments

Like most of the existing link prediction research, we tested our link prediction by applying our model to the first chronological timeframe of the data set and seeing how well it performs predicting links in the second timeframe. The accuracy is equal to the number of correctly guessed links divided by the number of actual links in the training period. Murata used 0.5 as his training fraction, however, since our model is more complex, we expect that our training interval will need to be longer. Thus, we tested the Murata forumla on our network with various training fractions and compared those to the results to see how much training data we should use.

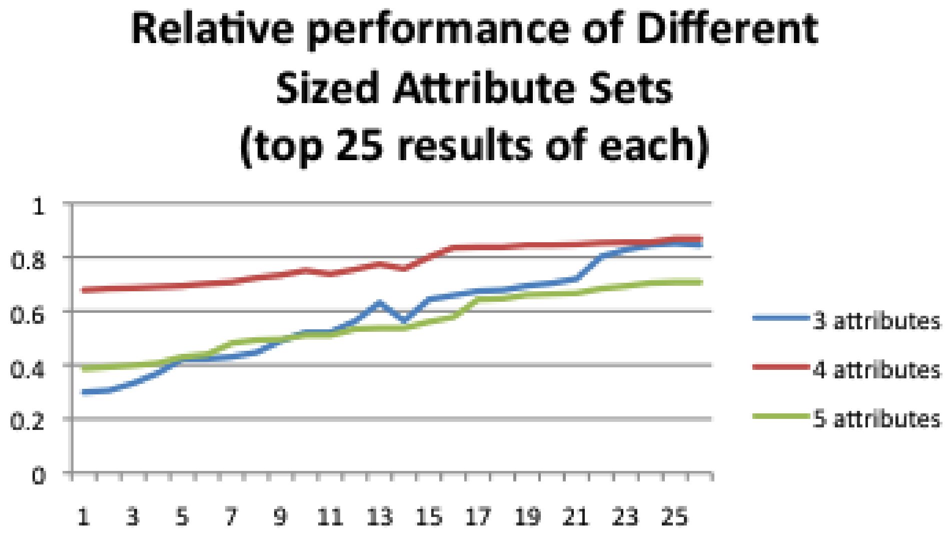


As expected, the more training data is used, the better the results. We opted to proceed with 0.7 as our training period for further analysis. The 0.49 accuracy made for a baseline for further comparison.

### 2.4.1 Testing Optimal Subset Size

In order to test the affect of set size on accuracy, we wanted to generate all powersets of the attributes and compare accuracies. However, there would simply be far too many subsets of the 162 original attributes to test. To decrease the number of trials run, instead of powersets of all of the attributes, we only included attributes that performed well. To do this, we first than the prediction algorithm on all 162 attributes and compared their results. We then chose the 25 best performing attributes to continue testing. See Appendix A.

Even then, the number of possible subsets of these 25 attributes took a prohibitively long time to run. To reduce the runtime, we used an approximated version of the original algorithm when large volumes of results were needed. Rather than all $n^2$ possible links between nodes during the training phase, we only used a random 20 percent of them. Although this decreased the accuracies drastically, the results could still be compared in terms of relative performance to each other.



The subsets of size 4 performed the best overall. The results illustrated that too few attributes does not add enough additional information the algorithm. The group membership then takes on too much weight in the scoring of links. Also considering too many attributes makes the results too general.

### 2.4.2 Normalization by Group Size

As it was found that the groups that aided most in predicting future interactions were also those with the greatest number of people, there was an attempt to normalize these results by group size. A list of the groups in order of size was compared to a list of the groups in order of predictive power. A measure for each group was derived from comparing the indices on the two lists, such that a group that was eighth largest but most predictive would have a score of 7, the difference between the indices. The results from this method were very difficult to interpret however, as groups that could be identified as “women’s sports” and were intuitively clusterable were found to have both very high and very low scores. (Girls’ varsity tennis was near the top, while eighth-grade girls’ basketball was near the bottom.) One thing to notice however is that this is corroborated by the results of McFarland’s “Bowling Young” paper however, as activities grouped as “sports” do not have consistent predictive power, as neither do activities grouped as “academic.” The complexity of the issues at hand was still interesting to see up close, and will act as a sobering lesson for future attempts at prediction of causal trends of this type. There are clearly other types of ways to normalize that may well prove more fruitful, but this was the only one attempted.

There was however a highly interesting and suggestive trend in the groups, if not one that had been looked for or anticipated. The five highest scoring (most predictive when normalized) groups contained one varsity sport, two junior varsity sports, a magazine editor, and the junior class board. The lowest were all activities from eighth grade. The lowest scores might also have been affected by the social shifts that occur between middle and high school, but the trend suggests that the amount of time or effort that must be dedicated to the activity in question drives its power as a link predictor more than the nature of the activity performed by the group. This would also fit will with intuition.

| Five Best Normalized Scores | | Five Worst Normalized Scores | |
|---|---|---|---|
| tennis_girls_var | 71 | band_8 | -121 |
| junior_class_brd | 70 | football_8 | -86 |
| passageway_editor | 57 | basketball_girls8 | -76 |
| basketball_boys_jv | 56 | basketball_boys8 | -75 |
| volleyball_jv | 55 | orchestra_8 | -70 |

2.5 Overall Performance

In order to then gauge how well the group attributes performed compared to the original Murata andAdamic/Adar formulas, we selected several subsets of attributes and ran them using the full data.

| Prediction Accuracies (0.7 training interval) | Prediction Accuracy |
|---|---|
| Adamic/Adar | 0.4318 |
| Weighted Murata | 0.4933 |
| Attribute Set : | |
| ['NHS', 'spanish_club', 'choir_concert', 'drunk_driving'] | 0.3773 |
| ['cauc', 'f', 'forensic_nat_for_league', 'orchestra_symph'] | 0.5180 |
| ['cauc', 'afam', 'spanish_nhs', 'band_march_symph'] | 0.5265 |

The benefits of adding attributes to the computations varies greatly. For some of the attributes, their inclusion significantly decreases the accuracy. However, for the best performing attributes, there are consistent improvements over the Adamic/Adar and weighted Murata methods. The results did not provide concrete enough evidence to be able to declare specific attributes to have generally great predicting power; however, the results show that group membership can be used to increase link prediction accuracy.

# 3. Conclusions

## 3.1 Results

Even while only using a very limited portion of the total data available, highly satisfying accuracies were achieved with the link prediction methods used. While the increases were modest, it was still interesting to corroborate and quantify the particular effects of group membership as compared with other phenomena within a community---work such as Putnam's focused primarily on the groups, and as such made it difficult to compare the magnitude of the effects within the total amount of interactions within the population.

While only suggestive, the findings when group impacts were normalized was very interesting, and inspiring of not only new analyses but new data sets entirely that could confirm or deny the correlation between predictive power and time or energy spent on a particular activity, and how this effect reacted with the activity itself.

## 3.2 Possible Future Work

A significantly greater amount of time and computing power could allow for a more thorough exploration of the group subsets, and would likely be the first step in a continuation of the project.

The most frustrating aspect of the findings however was the intense correlation between predictive power and group size, and the difficulty of normalizing these effects. This area would thus most likely be an active one in further work. One of the difficulties in interpretation has to do with the large spread in the abilities of groups within a certain cluster to predict social capital (as measured by other papers). Thus future attempts might take into account each group on an individual basis, as both previous work and the normalized results would support. This might also reveal more of a pattern within the already-used "point" method of scoring a group (the difference in index between its size and its predictive power) that were simply not visible when only the "types" of groups were taken in to account. Most interesting at present would be an attempt to quantify the engagement with a group, with the first step intuitively being the number of hours spent on the activity in question. However it is possible that new data would need to be collected to perform these tests, and he effort would likely be extenensive.

# Appendix.

A. Ranked list of individual attributes accuracies using attribute set prediction

| | |
|---|---|
| 1 | ('cauc',) |
| 2 | ('f',) |
| 3 | ('m',) |
| 4 | ('grade_10th',) |
| 5 | ('grade_12th',) |
| 6 | ('NHS',) |
| 7 | ('afam',) |
| 8 | ('grade_11th',) |
| 9 | ('pep_club',) |
| 10 | ('latin_club',) |
| 11 | ('choir_concert',) |
| 12 | ('spanish_club',) |
| 13 | ('spanish_club_high',) |
| 14 | ('key_club',) |
| 15 | ('drunk_driving',) |
| 16 | ('french_club_high',) |
| 17 | ('forensic_nat_for_league',) |
| 18 | ('Forensics',) |
| 19 | ('theatre_productions',) |
| 20 | ('band_march_symph',) |
| *21* | *('arts',)* |
| 22 | ('football_9',) |
| 23 | ('Debate',) |
| 24 | ('orchestra_symph',) |
| 25 | ('choir_treble',) |

Bibliography

[Liben-Nowell and Kleinberg, 2004] Liben-Nowell, D. and Kleinberg, J. The Link Prediction Problem for Social Networks.

[Loewi, 2010] Loewi, A. (2010). The unique importance and effects of personal interactions on the general exchange of information. unpublished.

[McFarland and Thomas, 2006] McFarland, D. and Thomas, R. (2006). Bowling young: How youth voluntary associations influence adult political participation. American Sociological Review, 71(3):401–425.

[Murata and Moriyasu, 2007] Murata, T. and Moriyasu, S. (2007) Link Predictions of Social Networks Based on Weighted Proximity Measures. IEEE/WIC/ACM International Conference on Web Intelligence.

[Putnam, 2000] Putnam, R. D. (2000). Bowling alone: the collapse and revival of American community.

[Putnam and Feldstein, 2004] Putnam, R. D. and Feldstein, L. (2004) Better together: restoring the American community.