VIII. Future Work

The relatively new and compelling field of Empirical Data Modeling (EDM) presents a significantly opportunity to improve the state of the art in semi-supervised machine learning. For example, the semi-supervised learning pipeline proposed by Glennan et al. was discussed in Section 2.5 and presented in Fig. 12. That pipeline represented a sequence of unsupervised clustering to produce a fully labeled data set, supervised classification, and then another unsupervised clustering phase to implement a fuzzy classification strategy. It is expected that some data may remain effectively unclassified (using the 'unknown' label) if very few or no labeled data are present in a particular cluster during either clustering step. Using EDM in the context of CPS network *semantic analysis* where the majority of data are expected to be *time series*, it may be possible to leverage causal relationships between clusters to classify more of the remaining unlabeled data. To demonstrate this, consider Fig. 53 which presents the output of a hypothetical semi-supervised data set labeling process.

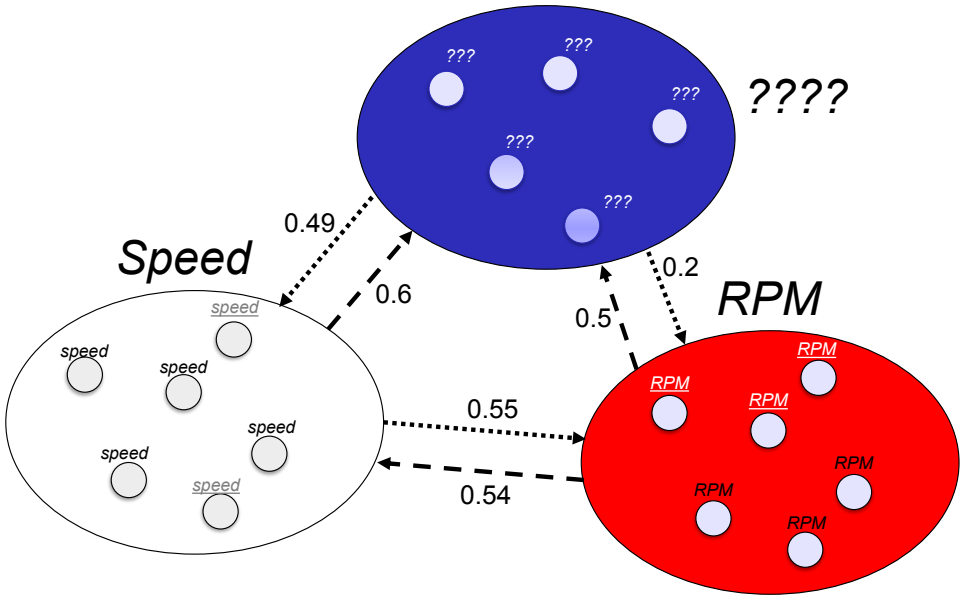

Figure 53. Example Result of Semi-Supervised Data Set Labeling

The points in the Speed and RPM clusters with black text represent the originally

labeled *time series*. In the case of automotive CAN network analysis, these labels may have come from correlation with limited J9179 diagnostic information. The other points in these two clusters with underlined grey or white text represent additional *time series* that were labeled by being in the same cluster of correlated *time series*. This process is referred to as *label propagation* in semi-supervised machine learning research [9, 10, 19, 80].

Current semi-supervised machine learning proposals have no reasonable means to label the points in the blue cluster. This is because they don't have the benefit of knowing causal relationships between clusters which is possible using EDM with *time series* data sets. The graph structure produced by adding causal links between clusters along with some limited manual reverse engineering may sufficient to train a supervised classifier to identify the blue cluster. For example, manual reverse engineering several vehicles may reveal that brake pressure *signals* consistently produce a cluster with the relative strengths of causal relationships to vehicle speed and engine RPM clusters as shown in Fig. 53[1]. These labeled weighted directed graphs could be used to train a supervised classifier to identify one or more of these clusters based on the graph structure and labeled clusters. A hypothetical result of this graph based cluster classification is shown in Fig. 54.

The weighted directed graph and labeled clusters might be iteratively passed as input to the classifier to label more clusters much like how the game of Sudoku is played. For example, imagine that there are several more unlabeled clusters than those shown in Fig. 54 with corresponding observed causal relationships between each pair of clusters. The classifier may not be confident in its classification of those other clusters using only the Speed and RPM cluster labels; however, it *is* confident in its

[1]The decision on exactly how the measure the strength of causation between *time series* in one cluster to those in another is very similar to selecting a *linkage* strategy during Agglomerative Hierarchical clustering

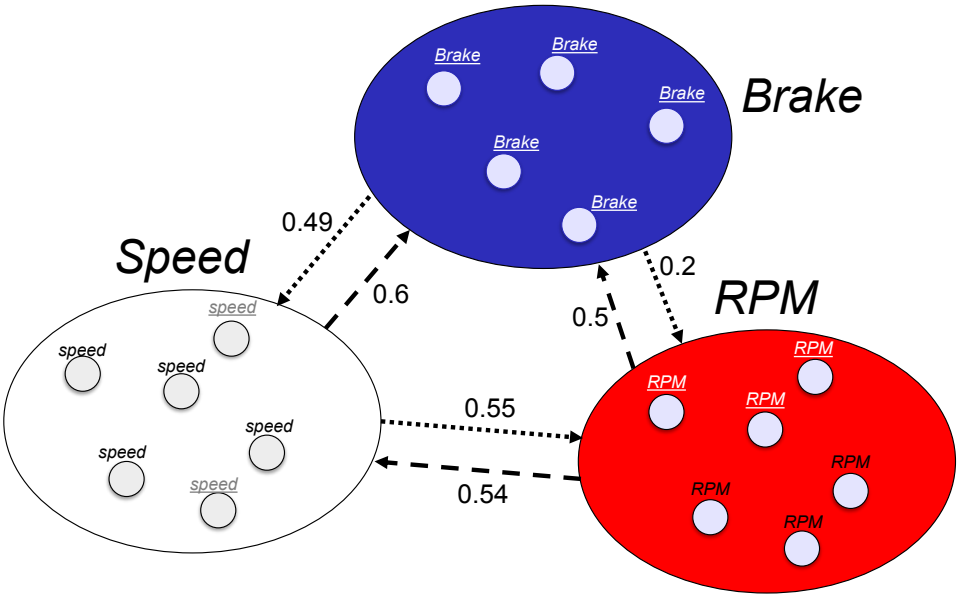

Figure 54. Example Result of Semi-Supervised Data Set Labeling Augmented With Graph Based Classification

classification of the brake cluster. The same graph is passed back to the classifier but now three clusters—Speed, RPM, and Brake—are labeled. Assume this third label increases the confidence for labeling one or more of the other unlabeled clusters of *time series* above a minimum threshold. This process is repeated until no new labels are assigned. At which point the researcher might decide to manually reverse engineer the remaining clusters and re-train the classifier. This process of iterative improving the graph based classifier might continue until the entire vehicle CAN bus can be accurately classified using the a few J1979 diagnostics and the weighted directed graph produced using the techniques described in this paper. Such a result would hopefully encourage manufacturers to abandon their policy of *security through obscurity* in favor of legitimate cyber-security mechanisms.