**Problem and Objectives**

Link prediction has proved to be useful in many real-world networks, such as recommending customer-centric products, and discovering novel drug-disease indications. In this project, we attempt to predict links in a simulated network graph using models/methods incorporating exogenous and/or endogenous factors.

**Simulated network graph**

An undirected graph *g* of 100 nodes was produced with seed = 1, such that for each *u < v*, (*u, v*) *∈ E* with the link probability between nodes *u* and v as *puv = yu yv* exp(-|*xu - xv*|), where *xu* and *yu* were generatedi.i.d. from Normal(0,1) and Uniform(0,1) respectively.The edges were independent. In this project, we would assume that the nodal attributes *xu* and *yu* of *g* are known but the true model is unknown.

We explored the graph’s connectivity and noticed that three nodes, namely nodes 28, 67 and 71 were disconnected from the main component (Figure 1). These nodes were removed from the graph network for subsequent analysis and modelling.

A picture containing outdoor object

Description automatically generated

Figure 1. Disconnected stimulated network graph

**Preliminary analysis**

Before fitting network models or applying predictive methods to the generated graph, exploratory analysis was performed and some key observations are detailed below. These paved the formulated hypotheses and motivations behind the proposed models and methods.

*Descriptive network characteristics*

Following the deletion of unconnected nodes, the new network graph yielded 97 nodes and 800 edges. Table 1 summarises some of the descriptive network characteristics.

Table 1. Summary of descriptive network characteristics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Nodes** | **Edges** | **Average Degree** | **Diameter** | **ASP** | **Density** | **GCC** | **ACC** |
| 97 | 800 | 16.49 | 4 | 1.99 | 0.17 | 0.31 | 0.31 |

*Preferential attachment*

We examined the nodal degree distribution and realized that it is rather homogenous. There is no apparent evidence of preferential attachment where nodes are linked to a few selected ones with high degrees. As such, we imagine that including k-star or geometrically weighted degree (GWD) variables in exponential random graph models (ERGM) or employing methods such as preferential attachment scoring may not enhance link prediction.

Chart, histogram

Description automatically generated

Figure 2. Distribution of nodal degree

*Transitivity*

With a global clustering coefficient (GCC) of 0.31, about 30% of the connected triples closed to form triangles. Therefore, the network has considerably high transitivity. Furthermore, its small average shortest path (ASP) of 1.99 and diameter of 4 highlights that this is likely an instance of the small-world network.

Such inherent cohesion in the network prompted us to incorporate triadic closure variables such as triangles and geometrically weighted edgewise shared partner (GWESP) in ERGM as well as utilise methods such as common neighbour and Jaccard scores favouring transitivity. The motivation behind these models or methods is the notion that the more common neighbours two nodes have, the higher the probability that a link would be formed between them.

*Homophily*

To assess the manner in which vertices of different degrees are linked together, we plotted the average neighbour degree against nodal degree. As seen in Figure 4, regardless of the nodal degree, we observe that the average neighbour degree remains relatively constant. This phenomenon is supported by the small positive assortative mixing by degree coefficient of 0.01, indicating that there is likely no evidence of nodes forming links with other nodes of similar degrees.

Chart, scatter chart

Description automatically generated

Figure 4. Average neighbour degree against nodal degree

Similarly, for *yu*, a relatively small positive assortative coefficient of 0.07 implies that there is weak assortative mixing with respect to *yu*. However, there is apparent strong assortative mixing by *xu*since its assortative coefficient is 0.36.This justifies the exploration of homophily variables in our models, particularly *xu* homophily, where nodes with similar *xu*have a higher chance of forming links.

*Nodal attributes*

We also preliminarily investigated the relationship between nodal attribute, *xu or yu* and degree using graph visualizations. Due to the extreme clutter in the plot of the entire network graph, subgraphs were created from the nodes with the maximum and median number of direct neighbours.

One positive finding is that nodes with larger *y* tend to form more links. For the yellow nodes with higher degrees, they are generally larger in size (larger *yu*) compared to the orange nodes with fewer degrees for both subgraphs (Figure 2).

A picture containing diagram

Description automatically generatedChart, radar chart

Description automatically generated

Figure 2. Relationship between *yu* and degreefor node with maximum and median neighbours

On the other hand, there does not appear to be any relationship between *xu* and degree since both the yellow and orange nodes appear to be evenly distributed in sizes (evenly distributed *xu*) (Figure 3).

A picture containing chart

Description automatically generatedRadar chart

Description automatically generated with medium confidence

Figure 3. Relationship between *xu* and degreefor node with maximum and median neighbours

Hence, we hypothesised that adding *yu* attribute baseline effect to the ERGM models would provide a better link prediction while *xu* attribute baseline effect would not.

**Link prediction methods/models**

We will transit from methods or models focusing solely on the network structure to those considering only nodal attributes before gelling these two sources of information in our models. Consequently, we can determine whether network structures or nodal attributes are more important in link prediction for our stimulated network.

To evaluate prediction performance, the top 10 node pairs with the highest scores or predicted link probabilities for each method/model were selected and the mean of their corresponding actual link probabilities was calculated. The higher the average actual link probabilities obtained, the better the method/model in link prediction.

*Prediction scores*

Prediction scoring methods largely makes use of the basic structural information that a network provides. In view of our preliminary analysis that the network is possibly transitive and unlikely to have preferential attachment, we used common neighbours, Jaccard and preferential attachment scoring to confirm our conjectures that the former two would perform better than the last method.

*ERGM without covariates*

Next, we considered ERGM models, with only variables taking into account the network structure, to serve as a comparison to the prediction scoring methods. We started out with a model using *edges*, *triangles* and *kstar(2)*. In a stepwise manner, we replaced *triangles* with *gwesp* and *kstar(2)* with *gwdegree* and optimised the respective gamma values for each model. Moreover, we also fitted a model with only *gwesp*, given that there is a likelihood of preferential attachment variables having no effect on link prediction based on preliminary analysis and the discrepancy between the coefficients’ signs of *kstar(2)* and *gwdegree* (Table 2). While the small positive coefficient of *kstar(2)* suggests possibility of preferential attachment in the network, the large negative coefficient for *gwdegree* hints

Table 2. Estimated coefficients of ERGM

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **ERGM without covariates** | | | | **Logistic models** | | **ERGM with covariates** | |
|  | **Model 1** | **Model 2** | **Model 3** | **Model 4** | **Model 5** | **Model 6** | **Model 7** | **Model 8** |
| **edges** | -3.300\*\*\* (0.168) | -4.264\*\*\*  (0.148) | 5.386\*\*\*  (0.501) | -2.740\*\*\*  (0.073) | -3.073\*\*\*  (0.171) | -3.021\*\*\*  (0.137) | -3.204\*\*\*  (0.184) | -3.109\*\*\*  (0.153) |
| **kstar2** | 0.028\*\*\*  (0.006) | 0.064\*\*\*  (0.004) | - | - | - | - | - | - |
| **gwdegree** | - | - | -5.023\*\*\*  (0.307) | - | - | - | 0.577  (0.608) | - |
| **triangle** | 0.159\*\*\* (0.020) | - | - | - | - | - | - | - |
| **gwesp** | - | 0.126\*\*\*  (0.035) | 0.107\*\*  (0.036) | 0.103\*\*\* (0.005) | - | - | 0.062  (0.040) | 0.049  (0.037) |
| **absdiff.x** | - | - | - | - | -1.252\*\*\*  (0.201) | -1.161\*\*\*  (0.078) | -1.144\*\*\*  (0.080) | -1.139\*\*\*  (0.080) |
| **absdiff.y** | - | - | - | - | 0.653  (0.703) | - | - | - |
| **absdiff2.x** | - | - | - | - | 0.045  (0.086) | - | - | - |
| **absdiff2.y** | - | - | - | - | -2.896\*\*  (0.914) | -2.027\*\*\*  (0.265) | -2.089\*\*\*  (0.275) | -2.021\*\*\*  (0.266) |
| **nodecov.x** | - | - | - | - | 0.066  (0.037) | - | - | - |
| **nodecov.y** | - | - | - | - | 2.711\*\*\*  (0.121) | 2.730\*\*\*  (0.121) | 2.692\*\*\*  (0.143) | 2.648\*\*\*  (0.135) |

\*\*p-value < 0.01, \*\*\*p-value < 0.00

at a predilection for lower degree nodes when new links are formed. In consideration of the ambivalence, it is essential to see how the removal of *gwdegree* may impact link prediction.

*Logistic models*

Logistic network model is a special case of ERGM, which enables us to include information of nodal attributes without any structural variables. Bearing in mind that there may be *xu* homophily and a relationship between *yu* and degree, we fitted a model which consists of the main effects of *xu* and *yu* using *nodemain*, as well as homophily effects of *xu* and *yu* using *absdiff* with both *power = 1* and *power = 2*. Thereafter, we conducted a stepwise backward elimination, removing each variable with the largest p-value one at a time until all variables are statistically significant at p < 0.05. Eventually, it culminated in a model with only the main effect of *yu*, *absdiff* of *xu* with *power = 1* and *absdiff* of *yu* with *power = 2* (Table 2).

*ERGM with covariates*

We would like to examine whether link prediction can be augmented with both structural and nodal attributes information. Each best-performing model from ERGM without covariates and logistic models were selected to yield a model with *edges*, *gwesp, gwdegree,* main effect of *yu*, *absdiff* of *xu* with *power = 1* and *absdiff* of *yu* with *power = 2* (Table 2). We also looked at the possibility without *gwdegree* owing to the previous ambiguous finding with regards to preferential attachment.

**Results and Discussion**

Table 3. Summary of model performance

|  |  |
| --- | --- |
| **Method/Model** | **Average link probabilities for top 10 node pairs for each method/model** |
| Common neighbours scoring | 0.518 |
| Jaccard scoring | 0.484 |
| Preferential attachment scoring | 0.481 |
| ERGM without covariates – Model 1 | 0.503 |
| ERGM without covariates – Model 2 | 0.540 |
| ERGM without covariates – Model 3 | 0.540 |
| ERGM without covariates – Model 4 | 0.514 |
| Logistic – Model 5 | 0.767 |
| Logistic – Model 6 | 0.776 |
| ERGM with covariates – Model 7 | 0.765 |
| ERGM with covariates – Model 8 | 0.777 |

*Methods/models with network structures only*

For methods/models focusing only on network structures (prediction scores and ERGM without covariates), their performances are somewhat similar.

Interestingly, preferential attachment scoring is not faring as poorly as anticipated, signaling potential preferential attachment in the network. Even though Jaccard scoring is based on the common neighbours scoring, it suffers a dip in link prediction accuracy, strengthening the possibility of preferential attachment since Jaccard scoring theoretically penalizes nodes with high degrees. While the positive coefficient of *kstar(2)* concur with the above phenomena (Model 1 and Model 2), it remains difficult to ascertain if such a characteristic indeed exists due to the contradiction arising from the large significant negative coefficient of *gwdegree* (Model 3) as well as the decreased accuracy in link prediction following the exclusion of *gwdegree* (Model 4).

*Models with nodal attributes*

Methods or models with only information on the network structures have poorer results than models involving nodal attributes (logistic models and ERGM with covariates). Moreover, even with inclusion of variables concerning the network structures, the best-performing model in ERGM with covariates (Model 8) only out-do the best logistic model very marginally (Model 6).

The insignificance of the *gwdegree* and *gwesp* terms (Model 7 and Model 8), after adjusting for the nodal attributes, proves that such structural information has little role in link prediction (Table 2). Additionally, the improved link prediction accuracy without *gwdegree* variable (Model 8) further affirms its redundancy and resolves the inconsistency from preferential attachment.

The former non-significant coefficient signifies nodes are not more prone to form links with other nodes with higher degrees than those with lower degrees. This agrees with our preliminary analysis that there is absence of preferential attachment. What differs is how the non-significant *gwesp* variable tells us that there is little inclination to form links between two nodes with many common neighbours, in contrast to the relatively high transitivity seen in the network. Relooking at the graph’s density of 0.17 and its GCC of only 0.31 may explain why transitivity may not be as crucial in explaining link prediction – the transitivity of the stimulated network is not particularly high.

In light of the above insignificant variables, we conclude that nodal attributes are more important than network structures in link prediction for our simulated network and it is critical that we analyse both aspects together to eliminate any potential confounding and erroneous deductions. As such, we choose Model 6 as our final model for link prediction.

The highly significant positive significant coefficient of *nodecov.y* means that larger *yu* is predisposed to more links while the highly negative coefficients of *absdiff.x* and *absdiff2.y* reveal that two nodes with smaller differences in *xu* and *yu* are more likely to form links (Table 2). For example, if |*xu* – *xv*| = 1, then the odds of *xu* and *xv* forming link is only 31% of the odds when *xu* and *xv* are equal. While the first two terms are expected as per preliminary analysis, the final term pointing to *yu* homophily is surprising.

**Conclusions**

Herein, we have provided evidence of how the humble logistic network model has provided reasonable link prediction accuracy and gained insights of which nodal attributes and homophily effects affect link formation between nodes. Nevertheless, more can be explored by trying out other models such as latent distance models, in view of how homophily is observed the network. Additionally, we can utilise different performance measures such as precision, recall and F-score to comprehend how the accuracy of link prediction is influenced by different metrics.