Network Regression Models

Benjamin Draves

May 2, 2018

Abstract

A central task of modern statistical network analysis is to uncover the probabilistic structure governing the behavior of the edges in a network. Several methods leverage the eigenstructure of the adjacency or Laplacian matrix to preform inference on the underlying probability model. In several applications, however, practitioners observe not only this network but additional statistics describing the vertices in the network. In this work, we look to appropriately incorporate this information to our estimation procedure through generalized linear regression models. Namely, we develop the *Edge Attribute Model* which incooperates vertex level information into the estimation of the edge random variables. We apply this model to both simulation and real-world connectomics data to better understand its feasibility in practice. Lastly, we offer a discussion which highlights possible improvements of the model developed here.

1 Introduction

In several modern statistical practices, interdependence of data is represented in networks or graphs. These objects encode a typically complex covariance relation not possible with other data structures. While this representation allows for a more complete depiction of the covariance structure of a system, it poses several statistical challenges. Everything from summary statistics to multi-graph models needs to be reconsidered or extended to fit this new class of data.

In this work, we focus on a regression problem in which we seek to identify edge attributes from vertex information in the $n = 1$ graph case. There has been a considerable amount of work done to construct a full distribution of random networks, entitled the Exponential Random Graph Model ([1], [2],[3], [4]). This model leverages the topology of the network by categorizing covariates based on network configurations which allow for a more complete representation of the information inherent in the network. As [5] notes, however, the estimation procedure for these types of models becomes regretably infeasible as the size of the vertex set increases past moderate sizes. Instead of using this powerful class of models, we instead choose to develop a more simple model that relies on more fundamental generalized linear model theory while imposing more restrictive conditions on the network in question.

This report is organized as follows; the development of the model we employ can be found in 2. A rigorous simulation experiment is completed in 3 to demonstrate the viability of the model developed in 2. In 4 we apply this model to real connectomics data in an attempt to better understand the connectivity structure of the mouse brain. Lastly we conclude and offer a discussion of possible modifications to the model presented here to extend to a broader class of data.

2 Edge Attribute Model

A central task in network statistics is identifying the underlying probability structure that governs the behavior of edges in the network. In several applications, a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ is observed where additional information about the vertices is known. For example, when a Facebook user's friendship network is observed, practitioners have knowledge not only of the user's friends, but also attributes describing their friend's interests and hobbies. From here, inferential questions about the user can be asked such as which shared interest is most likely to result in friendship.

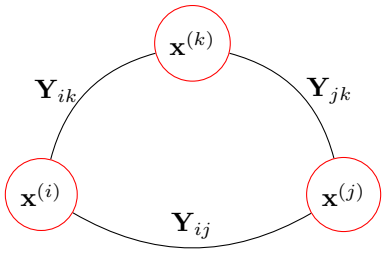

Figure 1: A simple network illustrating the edge attribute model. The random variables $\mathbf{Y}_{ij}$ describe the edge attributes while $\mathbf{x}^{(i)}$ describe the node attributes. A generalized linear model can be constructed to infer this joint relation.

Formalizing this concept, let $\mathbf{Y}_{ij} \in \mathbf{E}$ be the associated edge random variable between vertices i and j. Let $\mathbf{x}^{(i)}$ be covariate vector containing descriptive information about vertex $\mathbf{v}_i \in \mathbf{V}$. Our goal is to infer the relation between $(\mathbf{Y}_{ij}, \mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. This relation is described in Figure 1. One way in which we can infer this relation is to extend the generalized linear model to this network data. Suppose that $\mathbf{Y}_{ij} \sim \mathbf{F}$ a natural exponential family. For example, in the Facebook user model a reasonable model to consider is $\mathbf{Y}_{ij} \sim \text{Bern}(p_{ij})$ where $\mathbf{Y}_{ij} = 0$ corresponds to no friendship versus $\mathbf{Y}_{ij} = 1$ corresponding to friendship. As we will see later, by extending this model to general natural exponential family, we may use this same framework to analyze weighted networks. Then with link function $g(\cdot)$ we can model this relation by

$$g(\mathbf{Y}_{ij}) = \boldsymbol{\eta}_{ij} = \beta_0 + \sum_{\ell=1}^{p} \beta_j \mathbf{x}_\ell^{(i)} + \sum_{\ell=1}^{p} \beta_j \mathbf{x}_\ell^{(j)} + \sum_{k=1}^{q} \alpha_k h_k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \tag{1}$$

Here our parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be partitioned into two subvectors. $\boldsymbol{\beta}$ corresponds to the global effect of each covariate in $\mathbf{x}^{(i)}$. The $\boldsymbol{\alpha}$ parameters in this model correspond to the effect of *inter-nodal attributes* described by the data $h_k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. The data $h_k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ contain the information pertaining to similarity or correlation between the two vertices. For instance, returning to the Facebook user's example, the β_ℓ could correspond to the effect of the global level attributes such as how many friends the user has, their birth year, or the number of times they logged into Facebook. The $h_k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ data could for example be a grouping variable that is one if users i and j are in the same book club and zero otherwise. The α_k then measures the relative effect of this book club membership on shared Facebook friendship.

Notice that while in this example we took $\mathbf{Y}_{ij} \sim \text{Bern}(p_{ij})$ we could have easily allowed $\mathbf{Y}_{ij} \sim \text{Pois}(\lambda_{ij})$ or $\mathbf{Y}_{ij} \sim \Gamma(\nu, \frac{\mu_{ij}}{\nu})$. That is, in the model constructed here we have simply adapted the information inherent in the network to be modeled with known approaches such as generalized linear models and quasi-likelihood methods. One clear advantage of this construction is that we allow for the model given in (1) to be adapted to different data types in $\mathbf{Y}$ by simply updating our link $g(\cdot)$ as well as associated variance function $V(\mu_{ij})$ in the IRLS or quasi-likelihood procedures. Therefore, if we observe a weighted network $\mathbf{G}_\mathbf{w} = (\mathbf{V}_\mathbf{w}, \mathbf{E}_\mathbf{w})$ then we have flexibility in modeling the edge weights with an entire class of distributions. For example, if the $\mathbf{Y}_{ij}$ represent the number of messages sent in between users in a Facebook friend network the it may be reasonable to model this process using $\mathbf{Y}_{ij} \sim \text{Pois}(\lambda_{ij})$. If the edges are weighted by the time between last correspondence then it maybe reasonable to model $\mathbf{Y}_{ij} \sim \Gamma(\nu, \frac{\mu_{ij}}{\nu})$. Moreover, if a sample of networks $\{\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \ldots, \mathbf{G}^{(m)}\}$ are combined, rather naively, by averaging their adjacency matrices to form $\overline{\mathbf{A}} = m^{-1} \sum_{i=1}^{m} \mathbf{A}^{(i)}$ we may choose to model this process using a Gamma GLM.

While this model construction appears quite straight-forward, a more subtle point needs to be discussed. In standard generalized model theory we require that each response be independent. In several network applications, however, $\mathbf{Y}_{ij}$ need not be independent of $\mathbf{Y}_{jk}$. Indeed, one of the most attractive features of network data is the encoding of covariance structure within the object itself. While there are several applications where this requirement will be satisfied, it certainly will not be satisfied in all cases and therefore will need to be addressed. In the proceeding simulation study, we generate dependent data in an attempt to show the viability of this model even under this violated assumption through dispersion correction techniques. Some possible improvements to this model which address this obstacle are given in 5.

Having shown the flexibility of this model we now turn to apply these models to simulated data and real-world data. The simulated data demonstrate the ability of the model to be applied to the logistic regression and count regression settings. We then conclude by applying this model to the mouse brain connectome.

3 Simulation

3.1 Logistic Regression

In this section we show the ability of the edge attribute model to recover valuable information governing the probabilistic behavior of the edge set. We consider two examples; the first where $\mathbf{Y}_{ij} \sim \text{Bern}\left[\Phi\big(f(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\big)\right]$ and the second where $\mathbf{Y}_{ij} \sim \text{Pois}\left[f(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\right]$. As in either case the mean of $\mathbf{Y}_{ij}$ is simply the argument, we look to recover the functional form of f. In this simulation we consider the data $\mathbf{x}_i = (c_i, p_i)$ where $c_i \in \{1, 2\}$ is the class label for vertex $\mathbf{v}_i \in \mathbf{V}$ and p_i is an associated continuous covariate. In our first example, we define

$$f(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \begin{cases} -1 + \frac{1}{10}(p_i + p_j) & c_i = 1 = c_j \\ -2 + \frac{1}{10}(p_i + p_j) & c_i = 2 = c_j \\ \frac{1}{10}(p_i + p_j) & c_i \neq c_j \end{cases}$$

By defining $f(\cdot, \cdot)$ in this way, we see the each group has an inhomogenous effect on the edge random variable. Moreover, we see that the p_i covariates have a global effect on the edge presence. In this example, we let $n = 50$ with 20 vertices in group 1 and 30 vertices in group 2. Lastly, we generate the p_i from a Poisson random variable with rate parameter $\lambda = 4$. Having constructed $\mathbf{x}^{(i)}$, we generate a probability matrix $\mathbf{P}$ by $\mathbf{P}_{ij} = \Phi\left[f(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\right]$. Lastly, we generate a sample adjacency matrix $\mathbf{A}$ by sampling $\mathbf{A}_{ij} = \text{Bern}[\mathbf{P}_{ij}]$. Figure 2 shows the $\mathbf{P}$ matrix and associated sample $\mathbf{A}$. We note that as $f(\cdot, \cdot)$ decreases when i and j are in

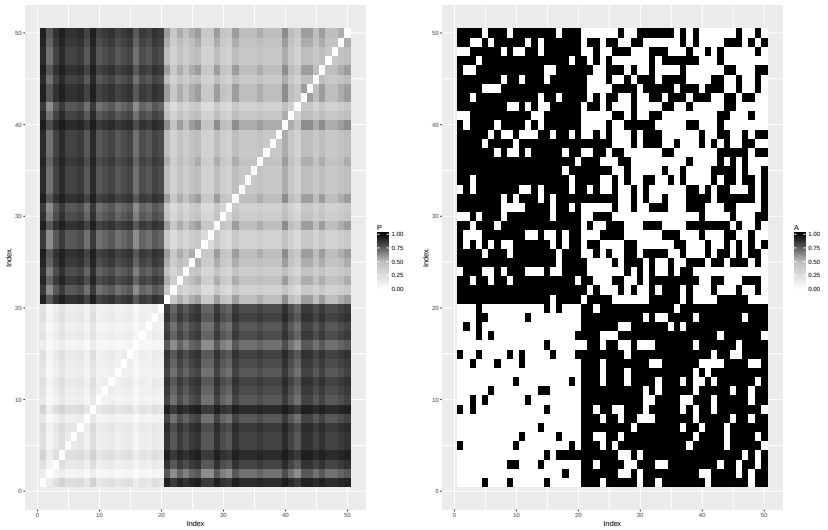

Figure 2: On the left the theoretical $\mathbf{P}$ matrix and on the right an elementwise sample $\mathbf{A}$ from $\mathbf{P}$. In practice we look to recover $\mathbf{P}$ from $\mathbf{A}$ using vertex specific information.

the same group we see that the off diagonal elements are considerably higher then the within group values. Moreover we see that while a considerable about the signal here is inherent in the group membership, the addition of the global effect of the continuous covariates p_i does indeed change the probability structure of this model. For this reason, we see that considering the global effects of each vertex covariate vector as well as the inter-nodal attributes discussed in 1 both contribute to the composition of $\mathbf{P}$. A visualization of the network associated with this adjacency matrix can be found in Figure 3. We now consider the problem of estimating the edge probabilities from the data given in each $\mathbf{x}^{(i)}$. One reasonable starting point would be to model the $\mathbf{Y}_{ij}$ using a Bernoulli regression model with canonical link using only grouping covariates. That

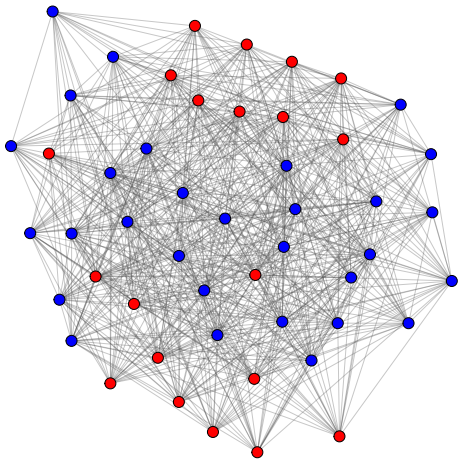

Figure 3: A realization of the network associated with the Bernoulli-edge model. In this model the edge model probabilities are functions of both group assignments as well as vertex attributes.

is, one initial choice is to model

$$\text{logit}\big[\mathbb{E}(\mathbf{Y}_{ij})\big] = \beta_0 + \beta_1 \mathbf{1}\{c_i = c_j\} \tag{2}$$

Implicit in this model structure is that the intra-group effects are identical across both classes[1]. Therefore, we expect this model to fit a single within class mean probability and a single inter-group mean. For this model we assume an undirected, no self loop structure. That is $\mathbf{Y}_{ij} = \mathbf{Y}_{ji}$ and $\mathbf{Y}_{ii} = 0$ for all $i, j = 1, 2 \ldots, n$. For this reason, we only look to model the elements of $\mathbf{A}_{ij}$ such that $j > i$. That is we preform a logistic-regression with response vector $\mathbf{Y}$ of size $n(n-1)/2$. The results of this model are summarized in Table 1. We first interpret that the coefficient estimates $\boldsymbol{\beta} = (\beta_0, \beta_1)$. The intercept term in this model, 1.418,

	Estimate	SE	p-value	Type	Deviance	Degrees of Freedom
Diff. Group	1.418	0.103	2e-16	Null Dev	1673.7	1224
Same Group	-2.056	0.133	2e-16	Res. Dev	1397.9	1223

Table 1: Associated model statistics for a homogoenous group model. While there is a clear difference between within group and inter-group structure the deviance appears high. This could be due to unmodeled variance in the response variable.

can be interpreted as follows; for two vertices in different classes, the odds they share an edge is 0.3492474. Moreover, the expected probability that two vertices in different classes share an edge is 0.8050247. The β_1 coefficient here represents the modulation in odds for two vertices in the same group to share an edge. Notice as $\beta_1 < 0$, we expect the odds that two vertices in the same class share an edge will decrease. Specifically, the odds two vertices in the same class share an edge is 0.5283481 and the expected probability these two vertices

[1]If the vertices in a network all belong to the same class, this model reduces to estimating a single probability of edge presence for the entire network. This estimation scheme is identical to estimating the parameter of the celebrated Erdos-Renyi model again showing the flexibility of the Edge-Attribute model.

share an edge is 0.3456988. While the associated test statistics suggest that these results are significant, we note that we may have evidence for overdispersion as the residual deviance is considerably larger than the associated degree of freedom in this model. A test for overdispersion found that $\hat{\sigma}^2 = 1.001635$ and failed to reject the null hypothesis that $\sigma^2 \leq 1$ and hence we conclude that these parameter estimates are indeed signficantly different than zero.

While this model may appear to the fit the data well, in light of Figure 2, we have cause to include additional covariates as $\mathbf{A}$ appears to demonstrate different behavior within each group. We fit the following model

$$\text{logit}\big[\mathbb{E}(\mathbf{Y}_{ij})\big] = \beta_0 + \beta_1 \mathbf{1}(c_i = 1 = c_j) + \beta_2 \mathbf{1}(c_i = 2 = c_j) \tag{3}$$

By constructing the model in this way, we allow for the inter-group relation to be modeled identically while accounting for inhomogenous group effects. This model is summarized in 2. Here, the baseline corresponds

	Estimate	SE	p-value
Diff	1.4178	0.1030	2e-16
Class 1	-3.3049	0.2381	2e-16
Class 2	-1.6627	0.2381	2e-16

Type	Deviance	Degrees of Freedom
Null Dev	1673.7	1224
Res. Dev	1336.6	1222

Table 2: Associated model statistics for an inhomogoenous group model. Here we see there is considerable difference between the probability structure of class 1 and class 2.

to a pair of vertices in different classes. For two vertices in different groups the expected odds they share an edge is 4.128205 with associated probability 0.805. The within group estimates then simply modulate these baseline odds. For instance, the odds two vertices in group 1 share an edge is given by 0.1515152 with associated probability 0.1315789. Similarly, the odds two vertices in group 2 share an edge is given by 0.7827869 with associate probability 0.4390805. A test for overdispersion as completed and we failed to reject the null hypothesis that $\sigma^2 \leq 1$. With this, we construct construct confidence intervals for these probabilities. These intervals are given in Table 3 Notice that as none of the confidence intervals intersect,

Group	95% Confidence Interval
Diff.	(0.77,0.83)
1	(0.09,0.19)
2	(0.39,0.49)

Table 3: 95% confidence interval table the estimated probabilities. As none of these intervals intersect, this suggests that each group probability is statistically different from one another and should be modeled as such.

this model suggests that each group structure is significantly different than the other. Moreover, this model reduces the AIC from 1401.921 to 1342.6 suggesting this inhomogoenous group model is a more consistent with the data.

Lastly, we include the continuous covariate present in each vertex. While there may be no interaction affect between p_i and p_j in this regression, their values may still provide meaningful information for modeling $\mathbf{Y}_{ij}$. For this reason, we wish to model their effect identically, while still including them in the mean function for this model. This is attained by modeling $\mathbf{Y}_{ij}$ by the following.

$$\text{logit}\big[\mathbb{E}(\mathbf{Y}_{ij})\big] = \beta_0 + \beta_1 \mathbf{1}(c_i = 1 = c_j) + \beta_2 \mathbf{1}(c_i = 2 = c_j) + \beta_3 (p_i + p_j) \tag{4}$$

By combining these covariates in this way we model their behavior jointly. In some applications it may be appropriate to take the difference, mean, or some other combination of these values. Moreover, for different datatypes (e.g. ordinal data), more complex combinations of these variables will be necessary. For simplicity, however, we only consider the sum of these variables here. The associated model fits are given in Table 4. The group covariates can be interpreted similarly as before. The third coefficient, however, has a different interpretation as $p_i + p_j$ is modeled as a continuous random variable. Explicitly, this model suggests that as $p_i + p_j$ increases by a single unit, we expect the log odds to increase 0.13429. Identically, for every unit increase in $p_i + p_j$ we expect the odds that an edge is present to increase by a factor of $\exp(0.13429) = 1.143724$. As

	Estimate	SE	p-value
Diff	0.25882	0.27567	0.348
Class 1	-3.40141	0.24309	2e-16
Class 2	-1.68111	0.14296	2e-16
Sum	0.13429	0.03039	9.94e-06

Type	Deviance	Degrees of Freedom
Null Dev	1673.7	1224
Res. Dev	1316.6	1221

Table 4: Associated model statistics for an inhomogoenous group model with global covariates included. There still is a considerable difference in group structure but also an effect due to the global covariates.

this quantity is greater than 1 we expect that as $p_i + p_j$ increases the probability of an edge connecting $\mathbf{v}_i$ and $\mathbf{v}_j$ to increase. A visualization of this model's fitted values are compared to the true probability matrix $\mathbf{P}$ in Figure 4. As evident by the figure, we see that this model fits the data very well. An analysis of deviance

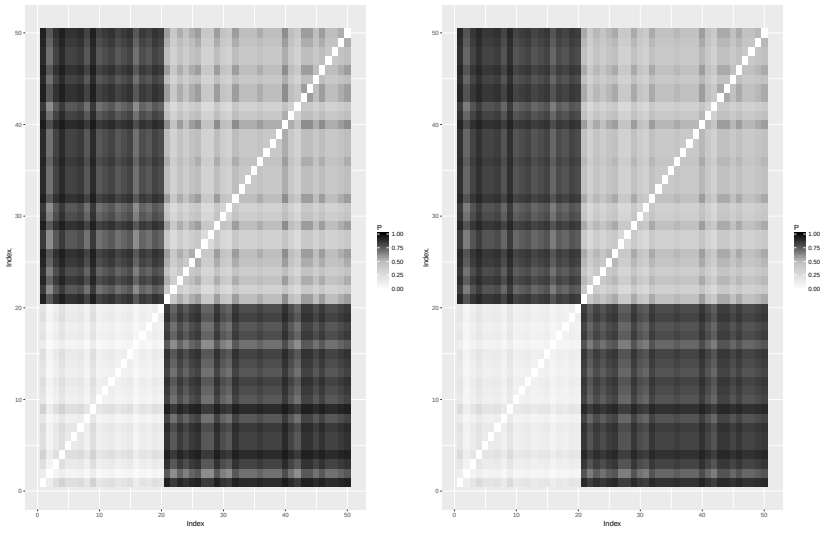

Figure 4: A comparison of the true probability matrix (left) and the estimated probability matrix (right). Here the edge attribute model does very well in recovering the probability structure of the edge variables.

was completed to compare all three models considered here. As expected, the homogenous model preformed the worse. The inhomogenous model provided significant improvement at the $p = 0.05$ level. Lastly, the inclusion of the continuous covariate improved on the inhomogenous group model at the $p = 0.05$ level.

This example shows the ability of the Edge Attribute model to recover the probability matrix in the unweighted network model. Moreover, we show that the inclusion of both global covariates as well as inter-nodal attributes can improve the estimation of the edge probabilities. We now turn to the weighted network model where we use count regression to infer properties of the edge random variables $\mathbf{Y}_{ij}$.

3.2 Count Regression

In several statistical applications the data is organized in a weighted-network where the edge weights correspond to the relative connectivity between vertices. In this way, the weighted network model describes not only statistical dependence but the strength of this dependence between subjects. Given two vertices $(\mathbf{v}_i, \mathbf{v}_j)$ with covariate vectors $(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ we seek to understand the connectivity between $\mathbf{v}_i$ and $\mathbf{v}_j$. One way to approach this problem is to use Edge Attribute Model where $\mathbf{Y}_{ij} \sim \text{Pois}\left[f(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\right]$. For this simulation, we let $n = 25$ and assign 10 vertices to group 1 and 15 vertices to group 2. Next, we sample p_i from a Poisson random variable with $\lambda = 2$. Lastly, we define f as follows

$$f(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \begin{cases} 10 + (p_i + p_j) & c_i = c_j \\ p_i + p_j & c_i \neq c_j \end{cases}$$

We organize these rate parameters in the matrix $\mathbf{L}$ where $\mathbf{L}_{ij} = f(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. From here we can generate our response random variables $\mathbf{Y}_{ij}$ by sampling from the distribution Pois($\mathbf{L}_{ij}$). The rate matrix $\mathbf{L}$ as well as sample weighted adjacency matrix $\mathbf{A}$ are displayed in Figure 6. As in Bernoulli model, our goal is to recover

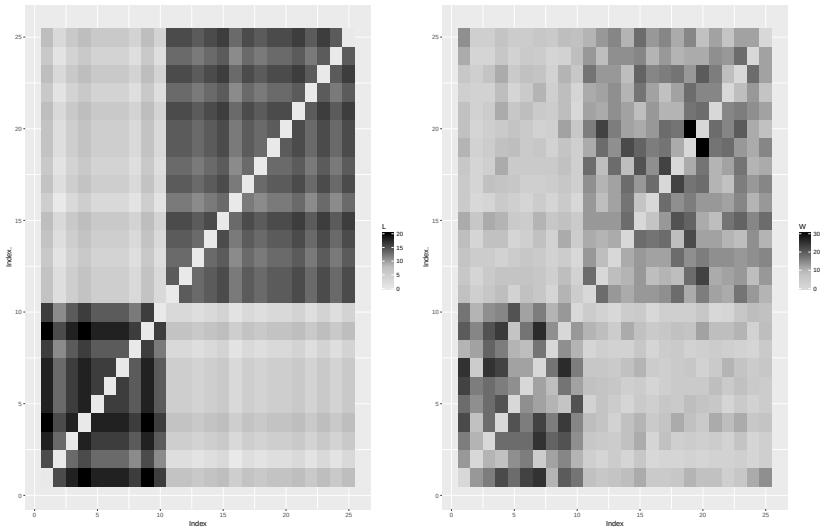

Figure 5: A comparison of the true rate matrix and the sample-weighted adjacency matrix (right). Here, as the rate also expresses the variance, the sample is much more variable than the Bernoulli model.

the functional form of f which characterizes the edge behavior in this model. We begin our inference by considering the model in which we fit a Poisson regression with continuous covariate as well as homogeneous group covariates. That is we fit

$$\log\left[\mathbb{E}(\mathbf{Y}_{ij})\right] = \beta_0 + \beta_1 \mathbf{1}(c_i = c_j) + \beta_2(p_i + p_j) \tag{5}$$

The resultant model statistics are given in Table 5. We notice here that both covariates are highly significant.

	Estimate	SE	p-value
Diff.Class	1.138188	0.059976	2e-16
Same.Class	1.058968	0.041874	2e-16
Sum	0.103805	0.009372	2e-16

Type	Deviance	Degrees of Freedom
Null Dev	1165.98	299
Res. Dev	374.65	297

Table 5: Associated model statistics for a homogeneous group model with all covariates included. There is a clear class effect as well as continuous covariate effect but it appears the model may be overdispersed.

Moreover, $\exp(\hat{\beta}_0) = 3.034692$ corresponds to the expected number of weights on an edge connecting two vertices in different classes with continuous covariates of value zero. Moreover, $\exp(\hat{\beta}_1) = 2.910137$ is the modulation in the mean number of weights on an edge connecting two vertices in the same class. That is, we expect 2.910137 times more weight on intra-class edges than on inter-class edges. Lastly, $\exp(\hat{\beta}_2) = 1.114683$ corresponds to the modulation in the baseline for every unit increase in $p_i + p_j$. That is for every unit increase in $p_i + p_j$, we expect to see 1.114683 times more weight on the edge connecting i and j. While this model appears to fit well, we note that there may be overdispersion based on the residual deviance. The estimates variance was given by $\hat{\sigma}^2 = 1.194257$. A test for overdispersion was conducted and resulted in a test statistic value of 354.6943 with critical threshold 338.193. With this, we reject the null hypothesis that $\sigma^2 \leq 1$ and therefore need to address overdispersion in this model. To account for this, we employ a Negative Binomial regression model with a log-link which inherently models underlying variability in the rate parameter. The model statistics for this GLM are given in Table 6.

A likelihood ratio test was completed to test if the negative binomial model did indeed improve the Poisson model given above. After completing the test, we see that the negative binomial model does statistical

	Estimate	SE	p-value
Diff.Class	1.11011	0.06507	2e-16
Same.Class	1.06820	0.04463	2e-16
Sum	0.10857	0.01039	2e-16

Type	Deviance	Degrees of Freedom
Null Dev	1017.07	299
Res. Dev	326.93	297

Table 6: Associated model statistics for a homogeneous group model with all covariates included using a Negative Binomial GLM. By considering this exponentially family, we successfully remove the overdispersion problem.

improve the model at the $p = .05$ level suggesting that this model is more consistent with the data. Moreover with an AIC reduction from 1010532 to 1541.169 we conclude that the negative binomial is the preferable model. A plot of the resulting model estimates are given in Figure 6. From this visual, it is apparent that

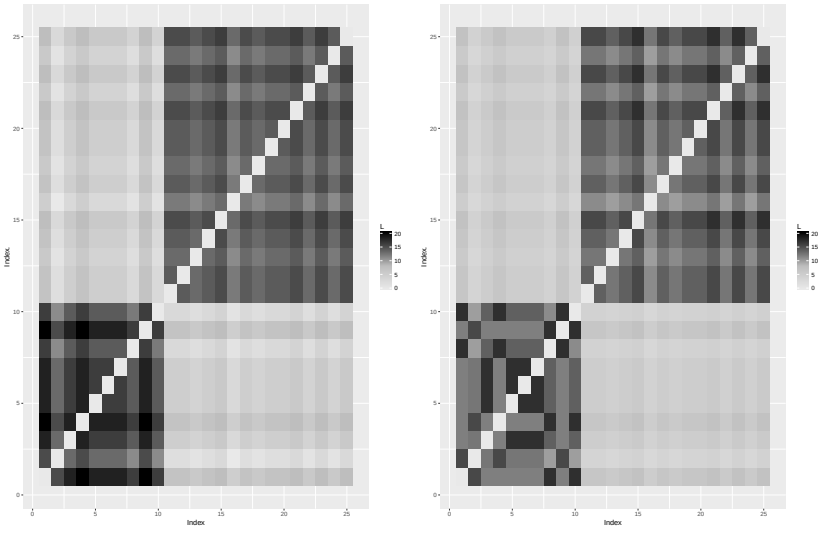

Figure 6: A comparison of the true rate matrix and the estimated rate matrix. Here, we again see the edge attribute model does very well at recovering $\mathbf{L}$.

the edge attribute model does in fact recover the structure of $\mathbf{L}$. We do note however, even in this simulation environment in which the functional form was known prior to modeling, the use of the Negative Binomial GLM was needed to address the variance structure of this model. That is, given the true underlying function generating $\mathbf{L}$, our model still required use of the Negative Binomial to account for the variance structure of this dataset. The reason for this is the effect of covariance among the responses. Here, we generated correlated data and used modeling methodology that assumed an independence structure. For this reason, we see an increase in the variance of our models and require modifications to address this obstacle. Therefore, in practice, we suggest using methods that address this altered variance structure explicitly, such as quasi-likelihood methods or use of different exponential families.

4 Mice Brain Connectome

In this section we apply the model developed in 2 to a mice brain connectome. The larger field of connectomics focuses primarily on the study of the way in which the nervous system is organized and connected. In particular a connectome is a collection of neural connections in any organ under investigation. For the sake of this application, these connections represent the number of nerve tracts passing through 332 labeled regions of interest. This data is organized in a weighted network with adjacency matrix $\mathbf{A}$. A visualization of this network is given in Figure 7. In addition to this data, each region of interest has an associated location within the brain. That is, each vertex in the network has a vector of class labels $\mathbf{x} = (c_0, c_1, \ldots, c_5)$ of varying resolution corresponding to a hierarchical class structure. The first corresponds to the hemisphere, the second

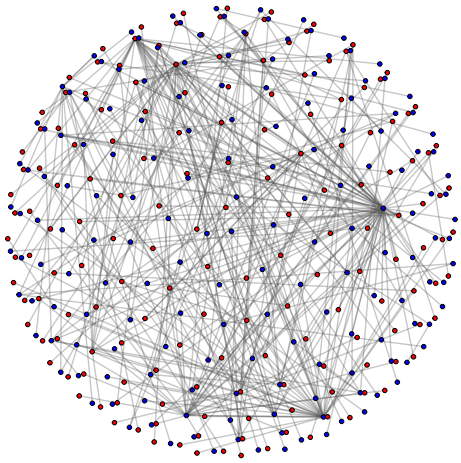

Figure 7: A visualization of the mouse brain connectome. Here we only plot the edge with the maximal tract number incident to each node in the network.

corresponds to a super structure such as the forebrain, and the remaining four class labels correspond to smaller nested substructures. We will refer to these smaller substructures as level 2, level 3, etc. We look to model the number of tracts connecting the two regions of interest as a function of the location in the brain in which each region resides. While we do not expect this spatial information alone will capture the complete behavior of the number of nerve tracts between regions of interest, we hope to uncover some statistical dependence between this value and the different superstructures that partition the brain.

To begin our modeling procedure, we begin with covariate design. As the hemisphere and the level 1 attributes (forebrain, midbrain, hindbrain, white matter, and ventricular system) partition the brain, we construct inhomogenous group variables for these superstructures. As the number of classes increase significantly for finer resolution classes, we choose to model these as homogeneous group effects. Moreover, as noted in 3, and several iterations of the model construction process, we choose to model this process using a negative-binomial GLM to account for possible overdiserpsion as well as covariance in the tract counts present in the connectome[2]. Moreover, for ease of interpretation, we choose to use the log-link in this application. From here, we build our response vector $\mathbf{Y}$ from the upper triangular matrix of $\mathbf{A}$ and construct our GLM using the covariates described above. The results of the model are summarized in Table 7. From this table we see that all variables are significant at the p = .05 level with the exception of the ventricular system which appears to have no effect on the number of tracts between two regions in this section of the brain. The baseline group here corresponds to two vertices in different hemispheres within different substructures. In this application, we see that the expected number of tracts between these two vertices is given by 801.961. We then partition the remaining covariates into two groups. The left hemisphere, right hemisphere, midbrain, white matter, level 4, and level 5 variables all have significantly negative values suggesting that *less* tracts exist between vertices in these regions. That is, for example, two vertices in the left brain are expected to have 0.9276346 times less tracts connecting the two vertices as compared to the

[2]We also note that the distribution of the edge weights is severely right skewed - the median is 84.0 with third quartile 503.8 and maximum value 95865.0. From this we have evidence to suggest that the variance does not grow linearly with the mean suggesting an alternative variance structure.

Covariate	Estimate	SE	p-value
Intercept	6.68706	0.01431	2e-16
L Hemi.	-0.07512	0.02182	0.000577
R Hemi.	-0.51989	0.02182	2e-16
Forebrain	0.09840	0.02989	0.000997
Midbrain	-0.28147	0.15384	0.067300
Hindbrain	0.38839	0.06513	2.47e-09
White Matter	-0.20445	0.07633	0.007393
Ventricular System	0.04547	0.54076	0.932982
Level2	0.09347	0.03461	0.006921
Level3	0.22387	0.05038	8.83e-06
Level4	-0.18563	0.05503	0.000744
Level5	-0.11979	0.04648	0.009953

Table 7: A table of statistics describing the model parameters of the Negative Binomial GLM with log-link. The majority of substructures of the brain have a significant impact on the number of tracts connecting the two regions of interest.

baseline group. The remaining covariates, forebrain, hindbrain, level 2, and level 3 all increase the baseline number of expected tracts as their associated β coefficient is greater than zero. With these group variables estimates, we conclude that it is generally very difficult to determine the exact connection between the brain regions and the number of tracts in the brain. In general, we see that the majority of tracts occur cross brain, while most intra-hemisphere connections are contained substructures of the level 2 resolution. We now turn to variance estimation to assess the consistency of this model.

Recall for the negative binomial model $\text{Var}(\mathbf{Y}_{ij}) = \boldsymbol{\mu}_{ij} + \frac{\boldsymbol{\mu}_{ij}^2}{\theta}$ where θ is the dispersion parameter. As $\theta \to \infty$, we see that this variance structure is identical to that of the Poisson. As we expect additional variance in this model due to the covariance among edges, we anticipate θ to be relatively small. A 95% confidence interval was constructed for the estimated $\hat{\theta} = 0.22966$ and is given by $(0.2273404, 0.2319876)$ suggesting overdisperson in this model. Using this estimate, we see that the estimated variance structure of this model is given by $\widehat{\text{Var}}(\mathbf{Y}_{ij}) = \boldsymbol{\mu}_{ij} + 4.354263\boldsymbol{\mu}_{ij}^2$. Here we see that there not only a significant quadratic term in the variance function but a dominating quadratic term suggesting that a Poisson regression would indeed be insufficient. Moreover, using quasi-likelihood methods may allow for more general approaches to modeling this variance. As stated above, however, it is difficult, if not impossible, in practice to differentiate variance attributable to noise versus the variance due to covariance of the response random variables. For this reason, we use the exponential family that most closely models this behavior.

To conclude, we see that the largest and smallest groups correspond to a reduction in the number of expected tracts while the groupings corresponding to the medium size groupings correspond to an increase in the number of expected tracts. One possible conclusion is that while most of the tracts in the connectome link the left and right side of the brain the majority of connections that occur within each hemisphere can be detected at the level 2 substructure resolution.

5 Conclusion

In this paper, we develop and analyze the *Edge Attribute Model* - a model that extends classical generalized linear model theory to the network regime. In this way, we leverage a well known analysis technique to aid in the analysis of the probabilistic structure of the edge set of a network. Using this simple model, we go onto show its estimation capabilities through extensive simulation study. Lastly, we apply this method to the mouse brain connectome to analyze the substructures of the brain. We find that while most tracts through the connectome occur across the brain, a rather simple partition of the brain at the level 2 resolution could be sufficient if we look to model local connectivity of the mice brain.

In this construction we noted that a rather restrictive assumption was made about the independence of the edge random variables. In some applications, the model presented here is sufficient and can be used as any other regression model. When the edge random variables are correlated however, more careful

analysis is needed. One possible model adjustment to address this issue is to model pairs of these random variables jointly. That is, form tuples of random variables $(\mathbf{Y}_{ij}, \mathbf{Y}_{jk})$ and use their joint vertex information $(\mathbf{x}^{(i)}, \mathbf{x}^{(k)}, \mathbf{x}^{(k)})$ and proceed with the same model used here. This model then of course relies on independence of these tuples and necessary adjustments will be required if this assumption is not satisfied. Another possible solution is to use the previously mentioned exponential random graph model. While computationally infeasible, these models explicitly address this issue by expanding the response space until the edge random variables are indeed independent. In either case, the computational cost as well as interprability of model parameters impose significant barriers to inference. In future works, an extension of this model to more accurately incorporate the covariance structure of the edge set is paramount. Only through modeling this structure can we truly extend the generalized linear model to the network regression problem.

Acknowledgements

I would like to thank Dr. Alexandra Badea for sharing the mouse connectome data analyzed here - it made for a very interesting project. I would like to thank Dr. Daniel Sussman in helping acquire this data. Finally, I would like to thank Liz Upton for providing valuable comments and advice on the computational procedures used here.

References

[1] Bruce A Desmarais and Skyler J Cranmer. Statistical inference for valued-edge networks: The generalized exponential random graph model. *PloS one*, 7(1):e30136, 2012.

[2] Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005.

[3] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

[4] Tom AB Snijders, Philippa E Pattison, Garry L Robins, and Mark S Handcock. New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153, 2006.

[5] Sourav Chatterjee, Persi Diaconis, et al. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461, 2013.

[6] John Ashworth Nelder and R Jacob Baker. *Generalized linear models*. Wiley Online Library, 1972.

[7] Eric D Kolaczyk. *Statistical analysis of network data: methods and models*. Springer Science & Business Media, 2009.

[8] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, volume 124. CRC press, 2016.