Analysing Attribute Based Multigraphs and Graph Prediction

Apratim Dey[*] and Diganta Mukherjee[#]
Indian Statistical Institute

Abstract

In this paper we study various relationships together using the multigraph structure, allowing for heterogeneity in the agent set. This is achieved using an attribute based network structure and analysis of the interlinkage that exists between the different layers of multigraph data and a prediction strategy for such graphs. The results are noteworthy on several counts. We observe that knowledge about the target graph helps in better prediction of the actual network rather than a blind prediction. In terms of sampling strategy, conditional on a fixed percentage, random sampling seems to work slightly better than other sampling strategies. This may be attributed to getting better information from "variety of data" rather than "localized or very specific data".

Keywords: Multigraph, Attribute based network, Sampling, Regression

AMS Classification: 05C82

[*]apratimdey.isical@gmail.com
[#] Corresponding author: diganta@isical.ac.in

1. Introduction

Like it or not, social networks are now fundamental elements of life. Meetings occur on a daily basis in school, at work etc. We are constantly considering adding new friends to our network or removing current ones. Friends and contacts are not only intermediaries in information diffusion, but they also influence consumption, voting or career decisions. Thus, it is of utmost importance to understand how social networks form. In this paper we study an extended problem, that of exploring whether one set of relationships among a given set of agents help in predicting the presence (or absence) of another relationship. The relevant definitions follow.

Networks, or graphs, are an important tool for representing relational data, that is, data on the existence, strength, and direction of relationships between interacting agents. Types of agents include individuals, firms, and countries. In its most basic form, a network consists of a set of n nodes and a set of edges, where nodes represent agents and edges the presence of a specific relationship between agents. The network can be represented by an $n \times n$ matrix $Y = (Y_{ij}), i, j = 1, \dots, n$; where Y_{ij} is a binary indicator, which takes the value 1 if an edge exists from i to j and is zero otherwise. By convention, $Y_{ii} = 0$. A pair of nodes is often called a dyad. Research on social networks is a well-established branch of study and many issues concerning social network analysis can be found in Wasserman and Faust (1994), Carrington et al. (2005), Butts (2008), Frank (2009), Kolaczyk (2009), Scott and Carrington (2011), Snijders (2011), and Robins (2013).

The interplay between strategic effects between nodes on a network and a network formation process was emphasized by Jackson and Rogers (2007), who extended the original preferential attachment model to create some links uniformly at random and some to friends of neighbours by preferential attachment. Social ties are created randomly (to family and relatives) and deterministically (to friends). We are interested in the network formation process and, more specifically, the strategic effects that precede link creation and deletion. Important works in this area are that of Konig et al. (2009), combining games on networks and network formation; and that of Jackson and Wolinsky (1996) and Bala and Goyal (2000), on the formation of networks itself.

1.1 Multigraph or Complex Graph

Human beings typically have several different types of relationships ranging from social and personal, to more business-like ones. These different types of relationships do not exist in isolation — one set of relationships of an individual can often be important for the other type of relationships. In simple terms a person's social relationships can also have a significant impact on her business outcomes (Joshi et. al. 2015). Social networks can have a strong influence on business networks. Belonging to certain clubs, being a member of specific groups in the real or virtual world, and attending particular events can play a key role in business affairs. At the least it has a reputational effect (to recall the well-known adage "One is known by the company one keeps") which can be useful in business deals. This link between social and business relationships has been well documented for the guanaxi in China (see Kali, 1999), the chaebols in Korea, and several communities like the Marwaris and Parsis in India (see Damodaran, 2008 for details).

As an example, consider a network with vertices representing different branches of an organisation. The edges apparent in such a network may then comprise of information, money, and personnel flows including cooperation, support, friendship and antagonism. These different edges should be considered simultaneously in order to understand the inter-organisational behaviour. Robins and

Pattison (2006) emphasise analysing these different edges jointly to understand social processes in the network and its implications for an organisation's performance. In an organisational network, it is also evident that different kinds of ties may appear within the same branch creating loops. For instance, friendships may be more common between individuals within a branch, which also indicate a higher propensity to turn to these friends for advice or support. (See Shafie, 2015 for details.)

This has even been studied in the context of networks by examining marriage alliances. Padget and Ansell (1993) examine marriages between 16 elite families in Florence in the early 1400s to explain the emergence of the Medici as an economic force. They show that in the marriage network more than half the paths relating the 16 families pass through the Medici making them twice as important as the family (the Guadagni family) that has the second largest number of such paths. Then Padget and Ansell (1993) go on to document that the Medici's rise in economic and political circles can be traced to their increasing importance in the marriage network, particularly under the astute leadership of Cosimo de Medici. The argument is that to the extent that marriage relationships were key to communicating information, brokering business deals and reaching political decisions, the Medici were much better positioned than other families.

A common definition of multigraphs (also called multiple networks) is graphs having several kinds of ties on the same vertex set (e.g. Robins 2013; Ranola et al. 2010; Koehly and Pattison 2005). Multigraph data structures can be observed directly and are common in contexts where several edges can be mapped on the same vertex pair, for instance social interactions of different kinds between a group of individuals (e.g. friends, colleagues, neighbours) or contact types (phone call, email, instant message) between and within departments of an organisation.

A complex graph according to Wasserman and Faust (1994) is: If a graph contains loops and/or any pairs of nodes is adjacent via more than one line the graph is complex. We must allow for multiple relations. This leads us to the study of multigraphs

1.2 Attribute based Network

The focal point of recent papers on network formation has been heterogeneity. Currarini et al. (2008) study link formation where individuals have types. They analyze the tendency of nodes to link to other nodes of their own types empirically. This so called homophily is also a key feature in the recent paper by Bramoulle and Rogers (2009). This heterogeneity may be easily incorporated in a network of graph by introducing a set of attributes that characterise each agent or node.

An Attribute based network is a graph in which the edges depend on certain properties of the vertices on which they are incident. In context of a social network, the existence of links between two individuals may depend on certain attributes of the two of them, for example their geographic location or socioeconomic status. We work with the underlying assumption that similar people connect to each other with higher probability. In the context of a social or a neural network, the connection between individual vertices depend on certain intrinsic qualities of the vertices themselves. It makes sense to consider the connection probabilities as a function of the vertex attributes. Sarkar et. al (2015) have shown that in the context of predictive modelling, attribute based networks are indeed worthwhile to study.

Plan of paper: In this paper we study various relationships together using the multigraph structure, allowing for heterogeneity in the agent set. This is achieved using an attribute based network structure. In the next section we describe the methodology and the data used for analysis. Sections 3

and 4 are devoted to the analysis of the interlinkage that exists between different layers of our multigraph data and the prediction problem for graph. Section 5 interprets the findings.

2. Model and Data:

ERGM: For multiple network studies, exponential random graph models (ERGMs) are particularly useful because they give evidence of how different networks relate to one another. Pattison and Wasserman (1999) present the exponential multigraph model (ERMM) which is based on hypotheses about the dependencies among multiple network tie variables. Applications of ERMMs can be found in Koehly and Pattison (2005) and Lazega and Pattison (1999).

The most commonly used model for a network is an exponential family model (Casella and Berger, 2002) of the form

$$Pr(Y = y;\ \theta) = exp\big(\eta(\theta)'\, Z(y) - \kappa(\theta)\big) \qquad (1)$$

These models are currently widely used for social networks (Strauss and Ikeda1990; Snijders 2002; Hunter and Handcock 2006). The first model of this type for social networks was proposed by Holland and Leinhardt (1981), and is known as the p1 model. See Suesse (2012) for details. Here θ is the vector of parameters whose values are exogenous to the network, e.g. characteristics of the agents. It may also include graph characteristics for any other graph(s) where these agents may be linked. This will become relevant for our multigraph setting.

Data used: We analyse social interactions and relationships through three social networks. In a New England law firm, USA, a study had been conducted on 71 lawyers in order to study the relationships amongst themselves, described in detail by Lazega and Pattison (1999) and Lazega (2001). The data are available online. The data include attributes of the lawyers, and three "directed" graphs, represented by their adjacency matrices. These graphs correspond to "Advice", "Friend" and "Work" networks among the lawyers. For example, if the (i, j)th entry in the adjacency matrix of "Advice" network is 1, then a directed edge exists from i to j, to be interpreted as i takes advice from j. Here, i and j are two lawyers. If the (i, j)th entry is 0, then no such edge exists. Each entry of the adjacency matrices is either 0 or 1. Similarly, the (i, j)th entry of "Friend" network decides if i considers j as a friend, and the (i, j)th entry of "Work" network decides if i works with j.
There are 8 attributes of the lawyers:

1. seniority
2. status (1 = partner; 2 = associate)
3. gender (1 = man; 2 = woman)
4. office (1 = Boston; 2 = Hartford; 3 = Providence)
5. years with the firm
6. age
7. practice (1 = litigation; 2 = corporate)
8. law school (1: Harvard, Yale; 2: UCon; 3: other)

3. Analysis of interlinkage using regression:

Let us define for $1 \leq i, j \leq 71$, and $1 \leq k \leq 3$, $y_{ij,k} = 1$ if a directed link exists from lawyer i to lawyer j in network k, and $y_{ij,k} = 0$ otherwise.

We denote "Advice" network by $k = 1$, "Friend" network by $k = 2$ and "Work" network by $k = 3$.

Define the following also for each i, j, k:
$Seniority_i$ = seniority of lawyer i
$Status_i$ = status of lawyer i
$Gender_i$ = gender of lawyer i
$Office_i$ = office location of lawyer i
$Years_i$ = years with the firm of lawyer i
Age_i = age of lawyer i
$Practice_i$ = practice type of lawyer i
$School_i$ = law school of lawyer i
$Degree_{i,k}$ = indegree of lawyer i in network k
$Degree_i = \sum_{k=1}^{3} Degree_{i,k}$ = indegree of lawyer i in all 3 networks combined
$Link_{ij,-k}$ = number of links from i to j in networks other than kth network

For analysis, we take recourse to the model (1), augmented with information related to the other graphs and some characteristics related to the present graph which are instrumental in the network evolution process.

It is intuitive that presence/absence of a directed link in a certain network, from lawyer i to lawyer j is dependent upon attributes of the two lawyers, their individual influence in that network (measured by the indegrees of the two lawyers in that network), total individual influence in all 3 networks (measured by the sum of indegrees in all 3 networks) and number of directed links present from lawyer i to lawyer j in the two networks other than the current network.

However, probably all these variables are not significant for predicting the probability of a link to be present. For example, probably, type of law school does not play any role in establishing a link between lawyers. Therefore, we perform a multiple logistic linear regression for each of the three networks to understand this dependence more clearly.

We regress $y_{ij,k}$ on

$$Seniority_i, Status_i, Gender_i, Office_i, Years_i, Age_i, Practice_i, School_i,$$
$$Seniority_j, Status_j, Gender_j, Office_j, Years_j, Age_j, Practice_j, School_j,$$
$$Degree_{i,k}, Degree_{j,k}, Degree_i, Degree_j, Link_{ij,-k}$$

Since we understand that some of the variables may be significant not because they have a direct influence on the links but because they are highly correlated with other variables that have influence on the links, we perform regression by variable selection also to see the difference.

3.1. For network "Advice":

After variable selection, the significant variables were
$Status_i, Gender_i, Office_i, Years_i, Age_i, Seniority_j, Degree_{i,k}, Degree_{j,k}, Degree_i,$
$Degree_j, Link_{ij,-1}$.

The last 5 variables are found to be significant, which we also expect, since they determine the influence of i and j in the networks and more influential people tend to connect with each other. Office location and seniority should also play a role in seeking advice as they determine convenience and wisdom respectively. The final regression model is as follows:

Coefficients:	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-1.31983	0.559929	-2.357	0.01842	*
Seniority_i	-0.3707	0.163249	-2.271	0.02316	*
Gender_i	-0.17484	0.118148	-1.48	0.13892	
Office_i	-0.25736	0.094221	-2.731	0.00631	**
Years_i	-0.02823	0.010839	-2.605	0.00919	**
Age_i	-0.01749	0.008186	-2.137	0.0326	*
Seniority_i	-0.00596	0.003106	-1.92	0.0549	.
Degree_(i,1)	-0.07163	0.0173	-4.14	3.47E-05	***
Degree_(j,1)	0.173572	0.015629	11.106	< 0.0000000000000002	***
Degree_i	0.033102	0.008462	3.912	9.17E-05	***
Degree_j	-0.04634	0.007182	-6.452	1.10E-10	***
Link_(ij,-1)	2.364726	0.088711	26.657	< 0.0000000000000002	***

Signif. codes:	'***'	0.001	'**'	0.01	'*' 0.05
(Dispersion parameter for binomial family taken to be 1)					
Null deviance:	4705.6	on	5040	degrees of freedom	
Residual deviance:	3083.3	on	5029	degrees of freedom	
AIC:	3107.3				

3.2. For network "Friend":

We see that the significant variables after variable selection are $Seniority_i, Status_i, Gender_i, Age_i, Practice_i, Degree_{i,k}, Degree_{j,k}, Degree_i, Degree_j$, and $Link_{ij,-2}$.

This is again expected. The last 5 variables are significant as expected. Friendships often are determined by age, gender and type of law practice. They denote similar mind set thereby causing better friendship relations. The final regression model is as follows:

Coefficients:	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-4.81833	0.631493	-7.63	2.35E-14	***
Seniority_i	0.024745	0.007535	3.284	0.00102	**
Status_i	-0.6966	0.216455	-3.218	0.00129	**
Gender_i	-0.38624	0.139551	-2.768	0.00565	**
Age_i	0.018822	0.009524	1.976	0.04812	*

Practice_i	0.28736	0.108012	2.66	0.0078	**
Degree_(i,2)	0.141633	0.017197	8.236	< 0.0000000000000002	***
Degree_(j,2)	0.209128	0.016505	12.67	< 0.0000000000000002	***
Degree_i	-0.01888	0.006107	-3.092	0.00199	**
Degree_j	-0.03805	0.005334	-7.133	9.83E-13	***
Link_(ij,-2)	1.532675	0.072115	21.253	< 0.0000000000000002	***

Signif. codes:	'***'	0.001	'**'	0.01	'*' 0.05
Null deviance:	3578.4	on	5040	degrees of freedom	
Residual deviance:	2670.9	on	5030	degrees of freedom	
AIC:	2692.9				

3.3. For network "Work":

After variable selection, the significant variables are
$Seniority_i, Status_i, Status_j, Degree_{i,k}, Degree_{j,k}, Degree_i, Degree_j, Link_{ij,-3}$

Partners tend to work with partners and associates tend to work with associates due to increased responsibility in the firm, so the dependence of link with status is understood. Of course, as usual, the last 5 variables are significant, as expected, since people tend to work with more popular (and more influential) people. The final regression model is as follows:

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.63172	0.347344	-13.335	< 0.0000000000000002	***
Seniority_i	-0.01004	0.004833	-2.078	0.0377	*
Status_i	0.281347	0.189507	1.485	0.1376	
Status_j	0.161539	0.110775	1.458	0.1448	
Degree_(i,3)	0.113018	0.012532	9.019	< 0.0000000000000002	***
Degree_(j,3)	0.158911	0.012824	12.391	< 0.0000000000000002	***
Degree_i	-0.00821	0.004872	-1.685	0.0919	.
Degree_j	-0.02944	0.00508	-5.795	6.82E-09	***
Link_(ij,-3)	1.435296	0.066968	21.433	< 0.0000000000000002	***

Signif. codes:	'***'	0.001	'**'	0.01	'*' 0.05
(Dispersion parameter for binomial family taken to be 1)					
Null deviance:	4261.2	on	5040	degrees of freedom	
Residual deviance:	3280.5	on	5032	degrees of freedom	
AIC:	3298.5				

Overall, comparing the proportional change in the deviance measure, the regression model for the "Advice" network performs the best. The other two regressions perform similarly to each other.

4. Prediction of "Work" graph:

Now suppose we are given information regarding the lawyers. Can we predict whether lawyer i and lawyer j have worked with each other?

We want to explore whether the other networks play any role in predicting the "Work" network. We assume the knowledge of networks "Advice" and "Friend". Work relationships are often established through advice and friendships among lawyers. Hence it is quite possible that "Work" network is better predictable given the other networks.

To do this prediction, we regress $y_{ij,k}$ on

$$Seniority_i, Status_i, Gender_i, Office_i, Years_i, Age_i, Practice_i, School_i,$$
$$Seniority_j, Status_j, Gender_j, Office_j, Years_j, Age_j, Practice_j, School_j,$$
$$Degree_{i,k}, Degree_{j,k}, Degree_i, Degree_j, Link_{ij,-k}$$

We note, now our $k = 1$ and $k = 2$, since $k = 3$ represents "Work" and that is assumed to be unknown.

We perform variable selection to remove variables that do not have a direct influence on the links to get the significant variables as
$Seniority_i, Status_i, Gender_i, Office_i, Years_i, Practice_i, Degree_{j,k}, Degree_i, Degree_j$,and $Link_{ij,-k}$

We observe an interesting thing here. In the earlier 3 regressions, the last 5 variables were significant. But in this case, after variable selection, the 17th variable i.e. $Degree_{i,k}$ is no longer significant. The final regression model is as follows:

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.07007	0.379184	-8.096	5.66E-16	***
Seniority_i	0.012498	0.004837	2.584	0.00976	**
Status_i	-0.47378	0.133323	-3.554	0.00038	***
Gender_i	-0.20248	0.083309	-2.43	0.01508	*
Office_i	-0.17903	0.072898	-2.456	0.01405	*
Years_i	-0.02096	0.008184	-2.561	0.01045	*
Practice_i	0.141114	0.069368	2.034	0.04192	*
Degree_(j,k)	0.175298	0.008206	21.363	<2.00E-16	***
Degree_i	0.019941	0.003573	5.581	2.39E-08	***
Degree_j	-0.05051	0.005166	-9.778	<2.00E-16	***
Link_(ij,-k)	2.548524	0.079235	32.164	<2.00E-16	***

Signif. codes:	'***'	0.001	'**'	0.01	'*' 0.05

(Dispersion parameter for binomial family taken to be 1)					
Null deviance:	8364.7	on	10081	degrees of freedom	
Residual deviance:	6267.2	on	10071	degrees of freedom	
AIC:	6289.2				

Now, using this regression equation and these estimates of regression coefficients, we predict "Work" graph. Let us call this prediction G^1. We then use G^1 as given and therefore have 3 known networks. Now using the tentative information from G^1 along with "Advice" and "Friend", we will predict "Work" again, calling this prediction G^2. Using G^2, "Advice" and "Friend" as given, we may predict "Work" again using this regression equation, and call this prediction G^3, and so on.

Thus, we will get a sequence of predicted graphs $G^1, G^2, G^3, \ldots$ that predict "Work". These graphs can be thought to form a stochastic process so they also give information about the evolution of "Work" network with time. Of course, with time, new links are created and old links are deleted, which reflect on "Work" graph. Within a small time span, the graphs in the sequence $G^1, G^2, G^3, \ldots$ will predict "Work", but long after, the predictions will not at all be correct as graph evolution has happened.

4.1. A measure of distance between graphs:

To understand the discrepancy between predicted graph and our actual "Work" graph, we use the measure

$$d(Pred, Actual) = \frac{1}{71^2}\sum_{i,j}\left(Pred_{ij} - Actual_{ij}\right)^2 \in [0,1] \qquad (2)$$

If this is near 0, we can say our predicted network and actual network are very similar. If this is near 1, there is a lot of discrepancy between reality and prediction.

Here we use the notation that for a graph/network A, a_{ij} denotes the probability of link from i to j.

Since we are using logistic regression, we can only get probabilities that $i \to j$ is linked so $Pred_{ij}$ will be probabilities, whereas $Actual_{ij} = 0$ or 1.

4.2. Prediction assuming no information about "Work":

In this case, since we do not assume anything about "Work", we reset $Degree_{j,3} = 0$ for all j (note $k = 3$ is the current graph). $Degree_i$ and $Degree_j$ are calculated based only on networks "Advice" and "Friend".

In the first iteration, we find that most of the link probabilities are very large, causing large discrepancy from actual "Work". In fact, $d(Pred, Actual) > 0.8$, which is a lot of deviation from the actual network. Further iterations do not improve this discrepancy much.

This clearly means that the $\widehat{y_{ij}}$ values (i.e. the predicted regression values) are very large, which causes the probabilities $\widehat{p_{ij}} = \frac{e^{\widehat{y_{ij}}}}{1+e^{\widehat{y_{ij}}}}$ to be near 1. So, we tune our regression model by a correction of

the intercept so that the average of $\widehat{p_{ij}}$ equals the actual graph density (Graph density is the ratio of actual number of edges in a graph to the total number of possible edges). We see that if -77.5 is added to the intercept obtained in regression, this is more or less satisfied.

Using this corrected intercept, we perform the regression, and we get G^1. However, we get G^2 using the original intercept obtained from regression. In both cases we find satisfactory discrepancy. However, in the second regression to obtain G^2 if we used the corrected intercept (i.e. original intercept – 77.5) then we get a large discrepancy between G^2 and our actual graph. Therefore, at the second round, we use estimated information about the "Work" graph and the estimate thus obtained in the second round is considered a reasonable candidate.

These regressions can be performed again so that the stochastic evolution of the graphs/networks can be observed, which are possible alternative realisations of our target graph. However, if we do have knowledge about some edges in "Work", we would like to use this information to predict the sequence $G^1, G^2, G^3, \dots$. Intuitively, if we have more information about our network, our predictions should be better.

4.3. Prediction with sampling according to seniority:

We perform this sampling using 20%, 40% and 60% of the existing edges and try to predict the actual graph. We order the lawyers according to seniority, and successively collect their links until we exhaust around 20% (or 40% or 60%) of the existing edges in "Work" is full. That is, we start from lawyer 1, collect his/her links, then go to lawyer 2, collect his/her links, and so on, till we exhaust 20% (or 40% or 60%) of the actual existing edges in "Work". Let us call this miniature incomplete version of "Work", G^0.

In predicting G^1, we therefore use our information of G^0. $Degree_{j,2}$ will now be the indegree of lawyer j in G^0, so not necessarily 0. $Degree_i$ will be the sum of indegrees of lawyer i in "Advice", "Friend" and G^0. The interesting feature here is, there is no need of intercept correction. Besides, at any stage, the discrepancies turn out to be smaller than that of prediction without any knowledge of "Work". Therefore, this reinforces our intuition that the more information we have, the better we predict.

4.4. Prediction with random sampling (without replacement):

First we collect all the existing edges in "Work" network and perform a simple random sampling without replacement to sample around 20% (or 40% or 60%) of these existing edges. A network/ graph G^0 is created with the same vertex set as "Work", but containing the sampled edges. No other edge is present in G^0. Then we perform regression using information from G^0 and the procedure is similar to that of the seniority sampling described above.

Here also one finds that there is no need of intercept correction. Again, the discrepancy at any stage is less than that of prediction without any information about "Work".

Another interesting observation is that, for a fixed percentage (20% or 40% or 60%), the first two iterations of regression by SRSWOR method yield **better** results than both Seniority sampling and prediction without any information. Further, as the percentage of known information increases, our

predictions also get better, in the sense of reduction in our distance measure (2). The results are tabulated as follows:

Type of prediction	Distance value after 1st regression	Distance value after 2nd regression
Prediction without any knowledge of "Work"	0.2294	0.1207
Seniority sampling (20%)	0.1187	0.1178
Seniority sampling (40%)	0.1165	0.1170
Seniority sampling (60%)	0.1137	0.1157
Random sampling (20%)	0.1183	0.1176
Random sampling (40%)	0.1144	0.1157
Random sampling (60%)	0.1124	0.1147

5. Interpretation of results

The results illustrated above are noteworthy on several counts. We observe that, on the whole, any knowledge about the target graph "Work" helps in better prediction of the actual network rather than a blind prediction. In terms of sampling strategy, conditional on a fixed percentage (20% or 40% or 60%), random sampling seems to work slightly better than seniority sampling. This may be attributed to getting better information from "variety of data" rather than "localized or very specific data". In random sampling, edges are obtained on a variety of lawyers unlike seniority sampling. This helps in better prediction.

When we are doing blind prediction, the first prediction is poor compared to the second prediction. But something very interesting happens at 20%. We see that for both seniority sampling at 20% and random sampling at 20%, the error in second prediction is smaller than that of first prediction. So performing regression twice actually gives us a better prediction of "Work" network in this case. However for both 40% and 60%, both seniority sampling and random sampling show the second prediction is worse than the first prediction. In fact, if one goes on doing the regressions, the errors increase with the number of predictions performed.

We see that at 25% (details not shown), however, the difference in discrepancy between the first and second regressions is very small, in fact, up to four decimal places, equal. Before 25%, the second regression performs better. After 25%, the first regression performs better. For final predictive purposes, the second prediction is taken rather than the first as the second prediction intuitively smoothens out discrepancies, like removing the need for intercept correction as in blind prediction.

References

Bala, V. and S. Goyal 2000. A Non-Cooperative Model of Network Formation.Econometrica68, 1181-230.

Bramoulle, Y. and B. W. Rogers. Diversity and Popularity in Social Networks". Working Paper, February 2009.

Butts, C. T., 2008. Social network analysis: a methodological introduction. Asian Journal of Social Psychology 11(1), 13–41.

Carrington, P., Scott, J., Wasserman, S. (Eds.), 2005. Models and Methods in Social Network Analysis, Cambridge University Press, New York, NY. Press.

Casella, G., and Berger, R. (2002), Statistical Inference (2nd ed.), Pacific Grove, CA: Thomson Learning.

Currarini, S., M. O. Jackson and P. Pin. An Economic Model of Friendship: Homophily, Minorities and Segregation". Econometrica, July 2009.

H. Damodaran, 2008. India's New Capitalists: Caste, Business and Industry in a Modern Nation, Palgrave McMillan.

Frank, O., 2009. Estimation and sampling in social network analysis. In: R. Meyers (Ed.), Encyclopedia of Complexity and Systems Science, Springer Verlag, New York, pp. 8213–8231.

Frank, O., 2011. Survey Sampling in Networks. In: P. J. Carrington, & , J. Scott (Eds.), The SAGE Handbook of Social Network Analysis, pp. 389–403, Sage.

Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs,"Journal of the American Statistical Association, 76, 33–50.

Hunter, D. R., and Handcock, M. S. (2006), "Inference in Curved Exponential Family Models for Networks," Journal of Computational and Graphical Statistics, 15, 565–583.

Jackson, M. O., and B. W. Rogers. "Meeting Strangers and Friends of Friends: How Random are Socially Generated Networks?" American Economic Review, vol. 97, No. 3, pp 890-915, June 2007.

Jackson, M.O. and A. Wolisnksy 1996. A Strategic Model of Social and Economic Networks, Journal of Economic Theory, 71, 44-74.

Sumit Joshi, Ahmed Saber Mahmud and Sudipta Sarangi (2015) Multigraph Network Formation with Strategic Complementarities, mimeo.

Kali, R. 1999. Endogeneous Business Networks, Journal of Law, Economics and Organization, 15, 615-636.

Koehly, L. M., Pattison, P., 2005. Random graph models for social networks: Multiple relations or multiple raters. In: P. Carrington, J. Scott, S. Wasserman (Eds.), Models and Methods in Social Network Analysis, Cambridge University Press, New York, NY, pp. 162–191.

Kolaczyk, E., 2009. Statistical Analysis of Network Data. Springer Verlag, New York.

Konig, M. D., C. J. Tessone and Y. Zenou. “Games of Dynamic Network Formation". February 2009.

Lazega, E. (2001), The Collegial Phenomenon: The Social Mechanism of Cooperation Among Peers in a Corporate Law Partnership, Oxford: Oxford University Press.

Lazega, E., Pattison, P. E., 1999. Multiplexity, generalized exchange and cooperation in organizations: a case study. Social Networks, 21(1), 67–90.

Padgett, J.F. and C.K. Ansell, 1993. Robust Action and the Rise of the Medici, 1400-1434. American Journal of Sociology, 98, 1259-1398.

Pattison, P., Wasserman, S., 1999. Logit models and logistic regressions for social networks: II. Multivariate relations. British Journal of Mathematical and Statistical Psychology, 52(2), 169–193.

Ranola, J. M., Ahn, S., Sehl, M., Smith, D. J., Lange, K., 2010. A Poisson model for random multigraphs. Fioinformatics, 26(16), 2004–2011.

Robins, G., Pattison, P., 2006. Multiple networks in organisations. Draft report.

Robins, G., 2013. A tutorial on methods for the modeling and analysis of social network data. Journal of Mathematical Psychology, 57, 261–274.

Koushiki Sarkar, Abhishek Ray and Diganta Mukherjee, Impact of Social Network on Financial Decisions, forthcoming in Studies in Microeconomics, 2015.

Thomas Suesse (2012) Marginalized Exponential Random Graph Models, Journal of Computational and Graphical Statistics, 21:4, 883-900, DOI: 10.1080/10618600.2012.694750

Scott, J., Carrington, P. (Eds.), 2011. Handbook of Social Network Analysis. Sage Publications, London.

Shafie, Termeh (2015) A Multigraph Approach to Social Network Analysis. Journal of Social Structure, 16, preceding p1-21. 22p.

Snijders, T. (2002), “Markov Chain Monte Carlo Estimation of Exponential Random Graph Models,” Journal of Social Structure, 1–40.

Snijders, T., Pattison, P., Robins, G., and Handcock, M. (2006), “New Specifications for Exponential Random Graph Models,”Sociological Methodology, 36, 99–153.

Snijders, T. A. B., 2011. Statistical models for social networks. Annual Review of Sociology 37, 131–153.

Strauss, D., and Ikeda, M. (1990), “Pseudolikelihood Eestimation for Social Networks”, Journal of the American Statistical Association, 85, 204–212.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press.