Statistical Methodology 7 (2010) 429–444

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Predicting profitability using advice branch bank networks

Avranil Sarkar[a], Stephen E. Fienberg[a,b,∗], David Krackhardt[c]

[a] *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA*

[b] *Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA*

[c] *Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA*

ARTICLE INFO

Article history:
Received 1 May 2009
Received in revised form
27 October 2009
Accepted 6 December 2009

Keywords:
Bootstrap
Centrality
Exponential random graph models
Hierarchical graph structure
Indegree
Outdegree
Regression models

ABSTRACT

The literature on social networks and their analysis has undergone explosive growth in the past decade. Network models have been used to study structures as diverse as the interaction of monks in a monastery, the links across the World Wide Web, and the structure of organizations. In much of this literature the network itself is viewed as the object of interest, and models are used to elucidate its structure. In this paper, we adopt a different perspective and we explore the role of network structure of organizations for prediction purposes. In particular, we work with data gathered on the advice-seeking habits of employees in 52 branches of a major North American bank corporation. We then use the network structure within each branch discovered via various exploratory analyses to predict the profitability of the individual branches.

1. Introduction

The past decade has seen an explosion of popular literature on social networks and their analyses, e.g., see the array of popular books on the topic [26,25,3,5,6]. Interest in networks has only intensified with the rapid growth of studies of the structure of the World Wide Web and the emergence of online social networks such as *Facebook, MySpace, Linkedin,* etc. This explosion has been paralleled by a burgeoning literature on network analysis, much of which is obsessed with the applicability of scale-free laws for degree distributions and other ad hoc methodology; e.g., see [2,4,8,21,9]. Yet there is a rich statistical literature on the analysis of social networks rooted in pioneering work by Holland and Leinhardt in the 1970s; e.g., see [17,10]. More recently, this literature has focused on the class of

∗ Corresponding author at: Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA.
E-mail addresses: avranils@stat.cmu.edu (A. Sarkar), fienberg@stat.cmu.edu (S.E. Fienberg), krack@cmu.edu (D. Krackhardt).

1572-3127/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2009.12.001

exponential random graph models (also known as p^* models) first proposed by Frank and Strauss [11] and extended by Strauss and Ikeda [23]. See also the more recent work of Hunter and Handcock [18]. The more recent literature has combined the older notion of stochastic blockmodels with clustering and/or mixed membership [1,16]. For a recent review of most of these statistical approaches, see [13]. One of the main stumbling blocks in the work on social networks has been the absence of asymptotics that allow an honest assessment of the goodness-of-fit of the models at hand for all but the simplest of models, such as the one examined by Frank and Strauss.

In this article, we adapt some of the ideas from the literature on exponential graph models to examine a collection of graphs for a collection of "parallel" entities; we then use the structure that emerges from an examination of the known social organization of the entities to predict a key quantity—profitability of the entity.

Every company has an organizational structure which is specified by the positions of the employees working in it and the kind of job that they have to do. How the employees work among themselves, however, cannot be understood just from the kind of job that they have to do; e.g., see the discussion in [7,24]. The performance of a company depends not only on the formal structure of the organization but also on the underlying structure of the way the employees work among themselves, something that we might glean from the informal structure of the organization. In this paper, we analyze the dependence of the performance of the company on the formal as well as the informal structure of the company. Much work has been done before on how these structures influence the individual performances of the employees but little work has focused on the influence on the performance of the organization as a whole. In this paper we use profit as a measure of performance at the organizational unit level and relate this profit to the unit's structure as a whole.

In Section 2 we describe the data, the graphical representation of the data, and the problem of interest. In Section 3, we focus on preliminary analyses of the data, including those utilizing network model structures, and in Section 4 we fit selected models to the data and assess their goodness-of-fit. We discuss the implication of the modeling exercise for the understanding of organizational structure in the final section.

2. Data

A number of years ago, one of the authors conducted a survey of all of the employees in 52 branches of a major North American bank. The questionnaire included the following "network-related" questions:

1. Whom do you talk to *at least once a week?*
2. Whom do you talk to *typically everyday?*
3. Whom do you go to for help or advice *at least once a week?*
4. Whom do you go to for help or advice *typically every day?*
5. Who comes to you for help or advice *at least once a week?*
6. Who comes to you for help or advice *typically every day?*

A primary goal of the study was to learn what "network features" and other structural aspects of relationships among the employees led to increased profitability of the individual branches. The dependent variable in this study, branch profit, was determined by the internal accounting department at the bank, which performs an annual review of the operations of each branch. One purpose of these accounting measures was to assess the performances of the branch managers during the year. A secondary purpose was to track trends of the contributions of the branches to the bank holding company's operational profits. Revenues were primarily comprised of loan interest payments, service fees, and bank transaction fees paid by branch customers. Expenses were primarily costs associated with personnel assigned to the branch, interest payments made to depositors, and branch operational costs such as utilities expenses. Fixed costs, such as rent and building maintenance, were not included in these measures. Profit for each branch was simply revenues minus expenses.

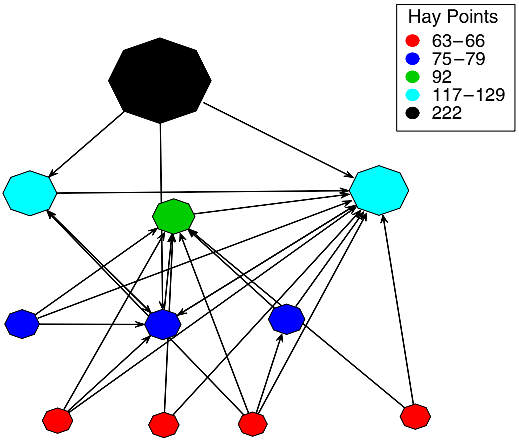

Fig. 1. Graph for a branch with 11 employees with the size of the nodes proportional to the number of Hay Points assigned to each of the employees' positions.

2.1. Network data representation

For each branch, we represent the questionnaire data in the form of a directed graph. The employees correspond to the nodes of the graph, and we include a (directed) edge from the *i*th node to the *j*th node if one of the following is true:

1. If employee *i goes* to employee *j* at least *once a day* and employee *j* confirms that employee *i comes* to him or her at least *once a week*.
2. If employee *i comes* to employee *j* at least *once a day* and employee *i* confirms that he or she *goes* to employee *j* at least *once a week*.

Thus, for the graph of each branch constructed in the above manner, if there is an edge from the *i*th to the *j*th employee, then employee *i* goes to employee *j* for advice. For our analysis, we consider only the advice questions, i.e., the last four questions. We illustrate the construction of the directed graph for a branch with 11 employees in Fig. 1. The sizes of the nodes are proportional to the Hay Points assigned to the employees. As we can see in Fig. 1, the employees can be classified into five different clusters or levels according to their Hay Points. These levels are represented with different colors. The lowest level has employees with Hay Points ranging from 63 to 66, the next level ones with Hay Points between 75 and 79, there is one employee with 92 Hay Points in the next level, and then there are two additional levels with employees whose positions have more than 100 Hay Points. The employee in the topmost level is the branch manager with 222 Hay Points.

2.2. Hay Points

The quantity labeled Hay Points in this study comes from a system of job evaluations created by the bank to provide pay brackets for different positions. Originally developed in the 1950s, the Hay Point system of job evaluations was developed by a large international consulting firm, the Hay Group.[1] Their system of job evaluations stems from a principle that different jobs require different skill and experience levels in order to carry out the responsibilities associated with each position. These Hay Points assigned to a job provide a salary range for the occupant. Adjustments to these salaries are made in accordance with individual performance levels; however, the range of these adjustments is restricted by the Hay Point guidelines.

[1] http://www.haygroup.com/.

Hay Points were originally developed to be generic descriptors, so that one could compare different jobs and salary levels not only within the firm but also across firms. For example, a Hay Point score of 200 in one firm should be equivalent to a Hay Point score of 200 in any other organization, and pay levels theoretically should also be comparable. In employing this system of job analysis and pay structures, however, some firms choose to specialize and tailor the descriptions to fit their own operations, culture, and organizational history. The bank company in the present study opted to adapt the generic Hay Point system by contracting with the Hay Group to develop a specific set of job evaluations for their own organization. Thus, while these specialized Hay Points are tied to pay scales within this bank, they are less comparable to ones for other firms, even other banks. Within the firm, similar position "titles" (such as Branch Manager, or Teller) often had different Hay Points associated with them. For example, a branch manager may be transferred to a new branch. On the surface, this may appear to be a lateral transfer. But if the transfer involves (as it usually does) moving to a larger branch with more responsibility, and the number of Hay Points associated with that new branch manager position is greater than the number of Hay Points associated with the prior position, this is considered a promotion. Tellers are promoted in steps defined by Hay Points as they take on more responsibilities, such as cashing larger checks without permission from a supervisor. Thus the number of Hay Points associated with a position is an indicator of both salary range and one's status within the bank.

The actual report that defined and determined the Hay Point evaluations is proprietary and was not provided to the researchers. However, the HR department of the bank did elaborate on the criteria used to derive the Hay Points for each position. These criteria included education level, specialized knowledge, supervisory responsibility for other persons (and their Hay Point levels), accountability (often phrased in terms of dollars that can be committed by the person on behalf of the bank), and amount of experience in different areas and at other Hay Point levels. The purpose in establishing the Hay Points system and applying it to each position was to systematize the pay scale for each job. Indeed, the Hay Points assigned to the positions in the bank are an approximate linear function (with a zero intercept) of the salary provided to the occupant within the position. Positions with twice the number of Hay Points are paid on average twice the salary. Every two years, these points are reviewed and adjustments are often made to reflect changes in the job requirements and in the competitive job market. Such adjustments are mostly made in the non-retail side of the bank; the retail branches involved in this study have not had significant Hay Point changes in more than eight years, although general pay scales associated with the Hay Points have increased over that time.

While the exact pay given to each employee was considered confidential, the Hay Points associated with the positions were "public" information within the firm and therefore were made available to the researchers. In the analyses that follow, we used these Hay Points of each position as an estimate of the salary and status level of the occupants of each position.

3. Preliminary analyses

3.1. Variability in the branches

Since we can consider each branch as a social network, one of the initial analyses that we implemented involved fitting some variations on p^* models that are commonly used for social network data. Our goal was to discover structural sources of differences across the branches before we attempt to predict profitability.

The branches of the bank have size ranging from 6 to 52 as Fig. 3 illustrates. Note that, for two branches of similar size, the distributions of the number of Hay Points of the employees in the branches are often quite different. In Fig. 4, we plot the distribution of Hay Points for branches of different sizes.

3.2. Exploiting the structure of the networks on the basis of the p^* model

Since each branch can be considered as a social network as explained in the last section, it is very natural to think that these networks can be well explained using a p^* model which is widely used for

Table 1
Summary statistics chosen in the p^* model.

Notation	Summary statistics
lev11	Edges among employees with Hay Point score 14.
lev22	Edges among employees with Hay Point score 80.
lev33	Edges among employees with Hay Point score 14.
lev12	Edges from Hay Point score 14 to 80.
lev23	Edges from Hay Point score 80 to 150.
lev34	Edges from Hay Point score 150 to 300.
lev21	Edges from Hay Point score 80 to 14.
lev32	Edges from Hay Point score 150 to 80.
lev43	Edges from Hay Point score 300 to 150.
dev	All the edges in the network not included above.

Table 2
Summary of results from regressing profit on the summary statistics.

Covariate	Estimate	Std. error	p-value
size	25.566	6.763	0.000458
lev11	−3.158	1.842	0.093286
lev34	77.581	27.729	0.007545
lev21	14.493	5.502	0.011529
dev	−11.725	3.788	0.003380
Adj. R^2	0.8723		

social network data. If we denote one realization of the network as a directed graph G, then a general p^* model for the probability of the graph is of the following form:

$$Pr(G) = C \cdot \exp\left[\sum_j \alpha_j T_j\right]$$

where T_j is a set of summary statistics for the network and c is the normalizing constant. In this case, we group the employees according to their Hay Points. Broadly speaking, for each branch there are four groups of employees with 14, 80, 150 and 300 Hay Points. The actual numbers of Hay Points often differ somewhat for different branches, but what we are really using here is the hierarchical structure of the network and not the actual value of the Hay Points. We choose as summary statistics the total number of directed edges from one cluster of employees to the other in the next upper level or the lower level in the hierarchy. These statistics account for how often one employee interacts with another one who is in the immediate lower or higher level in the hierarchy. Also, we use the total number of edges outside this hierarchical structure as a summary statistic. Table 1 gives the details of the summary statistics that we chose.

Now, if we assume that all the branches follow a p^* model with all these summary statistics, then the profit of these branches should also depend on these summary statistics. Thus we fit a linear model for profit vs. all of these summary statistics along with the size of the branches. Then we selected the best linear model using an exhaustive search and we estimated their coefficients in the linear model. We chose the best model as that with the minimum Bayesian Information Criterion (BIC). This model has as covariates size, lev11, lev33, lev21, lev43, dev. We summarize the results in Table 2 and the regression diagnostics in Fig. 2.

The negative value of dev represents the number of employees going for advice to people outside this hierarchical structure. This suggests that dev has a negative effect on the profit of the branches. Although the adjusted R^2 in Table 2 is 0.87, a value which we interpreted as very encouraging, the residual plot in Fig. 2 does not seem to be an independent observation from a normal distribution. One reason for this is the highly variable distribution of the employees in the different branches with respect to the Hay Points assigned to their positions.

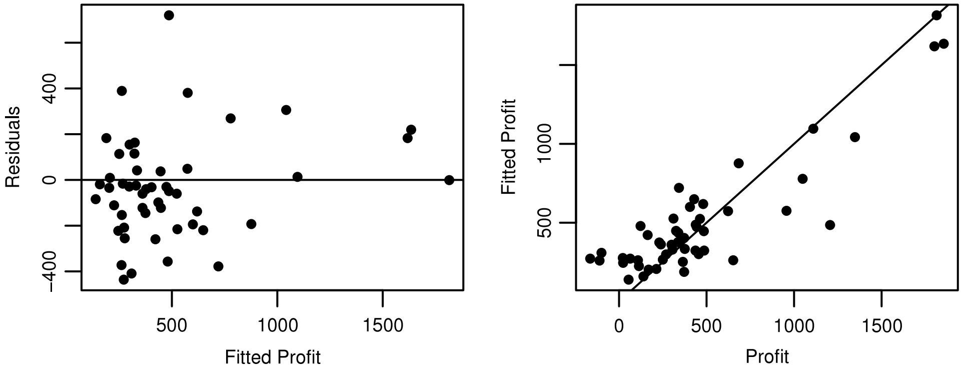

Fig. 2. Regression diagnostics for regressing profit on the summary statistics chosen in Table 2.

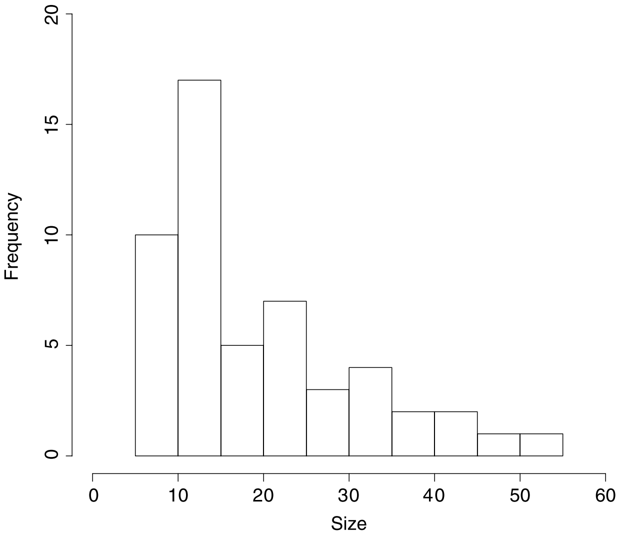

Fig. 3. Histogram for sizes of the branches.

3.3. Variability in the branches

The branches ranged in size from 6 to 52 as we can be seen in Fig. 3. Also, for two branches of similar size the Hay Points for the employees in the branches are not always similar. In Fig. 4, the distributions of Hay Points for branches of similar as well as different sizes are plotted. We can see in Fig. 4 that a large and a small branch have different Hay Point distributions; the same is also true for two small branches of similar size and two large branches of similar size. Because of this, the total numbers of edges between two different clusters of employees will be very different. In the data that we are using, if we focus on the bank branches after controlling these kinds of variabilities, we will have too few branches to do meaningful statistical analyses. Our emphasis in this paper is on the dependence of profit on the statistics relating to the structure of the network which are not affected too much by these kinds of variabilities. Since the profit of a branch depends on its size as well as the money that it invests in its employees (characterized by the Hay Points of its employees), however, we first need to work out the dependence of profit on these factors.

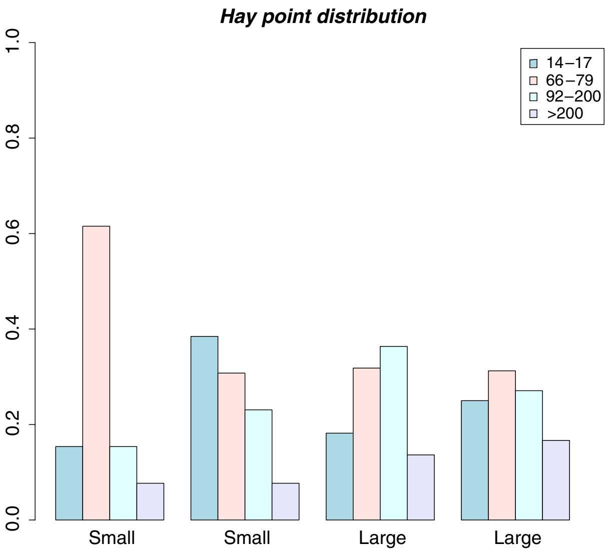

Fig. 4. Barplot for Hay Point distributions within branches with similar as well as different sizes.

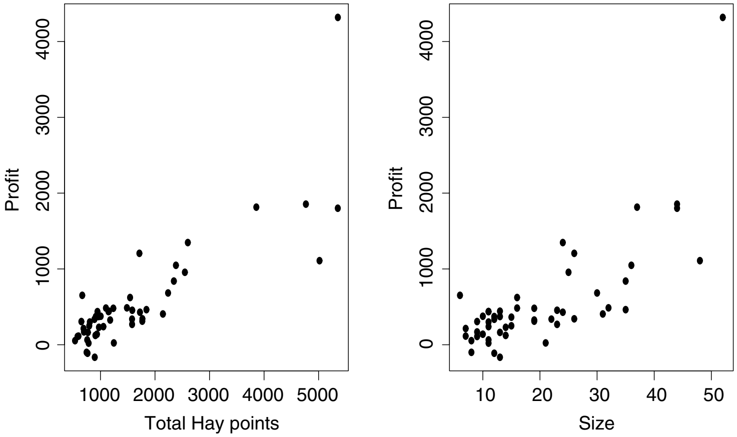

Fig. 5. Plot of profit vs. size and total number of Hay Points.

3.4. Variability of profit across branches

In Fig. 5, we plot profit against size and then against the total number of Hay Points for all the branches. We see that the randomness in the plot for profit vs. size is much more than that in the plot of profit vs. total number of Hay Points. It has been observed before that there is a lot of variability in the distribution of Hay Points even for branches of the same size. Thus, the kinds of jobs that the employees are doing for two branches of the same size appear to be different and hence the profits might well be different. Since our aim is to find some kind of association between profit of a branch and the structure of the graph representing the branch, it is important to eliminate the variability in profit due to size and Hay Point distribution. In order to do that, we will look at scaled profit for each branch where scaling is done with respect to the total number of Hay Points. In that way, we can adjust for both the size and the distribution of the number of Hay Points.

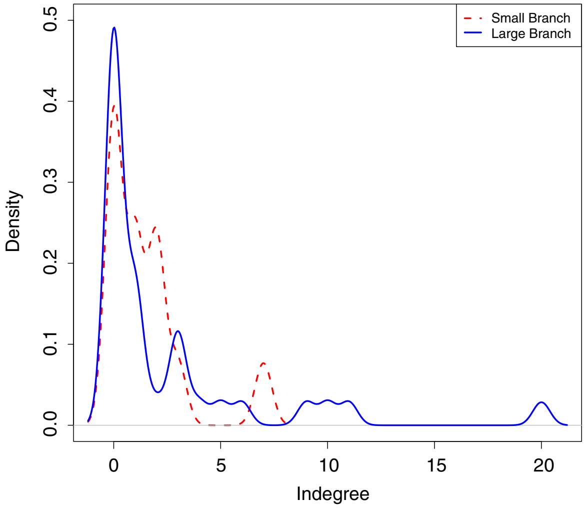

Fig. 6. Indegree distribution for a small and a large branch.

Table 3
Range of number of employees with low and high indegree for a small and a large branch.

Indegree	Small branch	Large branch
Low indegree	2–5	2–9
High indegree	5–17	11–29

3.5. *Analysis of network characteristics*

Earlier, we described the variables that vary across branch banks. Now we explore some characteristics of these networks which are very similar among all the branches.

3.5.1. *Common features in the network across different branches*

One of the most important features of a directed network is its indegree distribution. The indegree for a node in a directed graph is defined to be the number of edges received by that node from all other nodes in the graph. The indegree distribution is the distribution of indegrees for all the nodes in the graph.

In Fig. 6, we plot the indegree distributions for a small (size ≤ 23) and a large (size > 23) branch. As we can see, in both the branches there are two distinct groups of employees, the first with low indegree ($\leq$3 for small and $\leq$6 for large) and the second with high indegree ($\geq$6 for small and $\geq$9 for large). We observed this feature in most of the branches and illustrate it in the boxplot for number of employees with low and high indegree in Fig. 7. This means that in each branch there are some employees (the second group) to whom most of the other employees come for advice. For small branches, there are usually 2 to 5 employees with high indegree, with the median at 3, whereas for large branches the number of employees with *high* indegree varies from 2 to 9, with the median at 4. The number of employees with *low* indegree varies from 5 to 17 in the small branches with the median at 7.5 and 11 to 29 in the large branches with the median at 20. We summarize this information in Table 3.

In the next section we propose a measure for the branch networks based on these two groups of employees with low and high indegree.

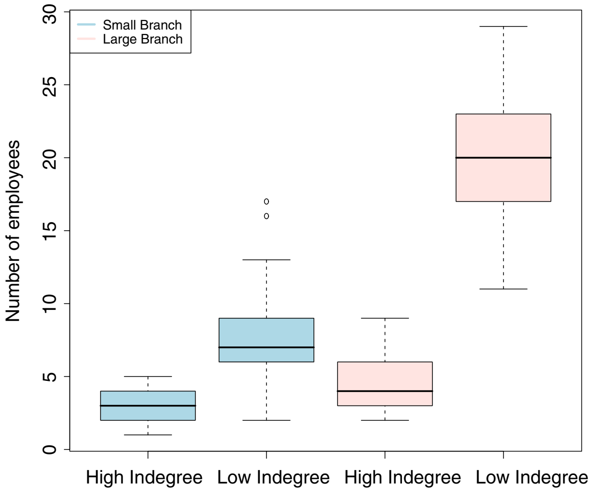

Fig. 7. Boxplot for the number of employees in both groups for small and large branches.

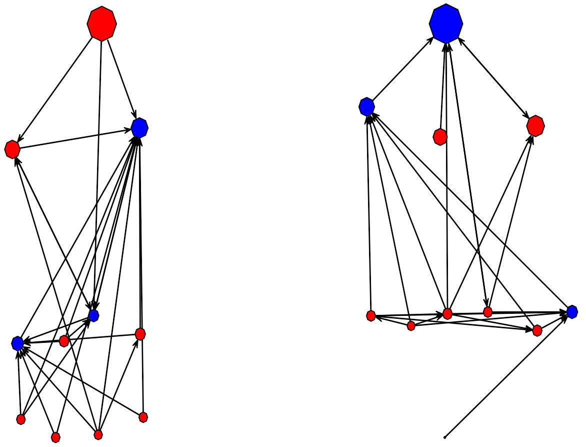

Fig. 8. Networks for high and low profit branches. The left network corresponds to a low profit branch and the right one corresponds to a high profit branch. The blue nodes represent employees with high indegree and the red nodes represent employees with low indegree. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. A new summary measure for the branch bank networks

4.1. Relationship between profit and some features of the network

To motivate our proposed measure, we provide the following illustration based on a comparison of typical high profit vs. low profit branches and note certain network characteristics that differentiate them. Let L be the group of employees in a branch with high indegree, which is often associated with informal leadership status (cf. [19]), and let S be the remaining group of employees in the branch.

In Fig. 8, we see the networks for two branches both of size 11—one with low profit and another one with high profit. The sizes of the nodes are proportional to the Hay Points of the employees. The blue nodes represent employees with high indegree (set L) and the red nodes represent employees with low indegree (set S). In both the branches, there are 3 employees with relatively high indegree

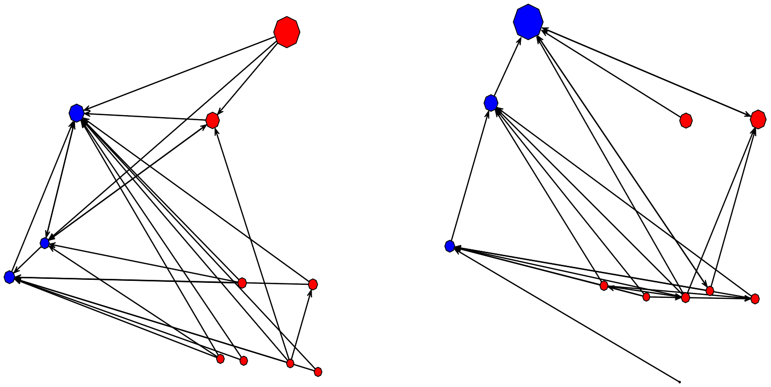

Fig. 9. Networks for high and low profit branches emphasizing the separate L and S employees. The left network corresponds to a low profit branch and the right one corresponds to a high profit branch. The blue nodes represent employees with high indegree (set L) and the red nodes represent employees with low indegree (set S). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

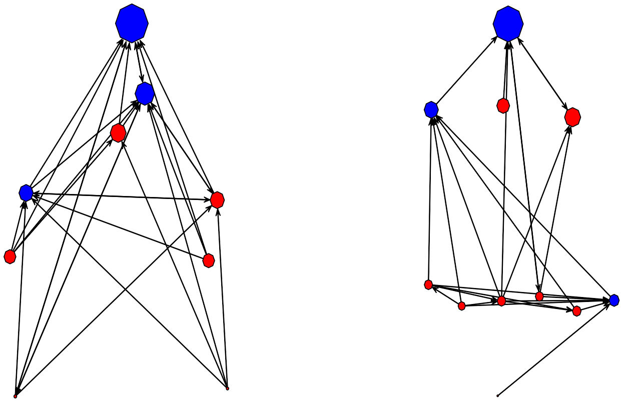

Fig. 10. Networks for high and low profit branches. The left network corresponds to a low profit branch and the right one corresponds to a high profit branch. The blue nodes represent employees with high indegree and the red nodes represent employees with low indegree. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and 8 employees with relatively low indegree. Fig. 9 shows the network for the same branches except that we have rearranged the nodes to emphasize the separate L employees and S employees.

The differences observed in the structure of the network for the two branches are listed below:

- The employees in L for the high profit branch come from different levels of the branch i.e., all of them differ significantly in their Hay Point designations, unlike the situation in the low profit branch where two of the employees have the same number of Hay Points.
- For each employee in group L, the employees in S coming to him or her for advice have more variability in their Hay Points in the high profit branch than in the low profit branch.

In Figs. 10 and 11, we plot the network for another low profit branch of size 9 and the same high profit branch from above. In this case, the groups of blue nodes for the two branches have similar properties, i.e., they come from different levels of the branch. The extent to which interactions among the red nodes bridge across varying levels, however, is far more pronounced in the low profit branch than in the high profit branch (see Fig. 11).

We summarize the features observed across all the low profit and high profit branches below. For any high profit and low profit branch at least one of the following is true:

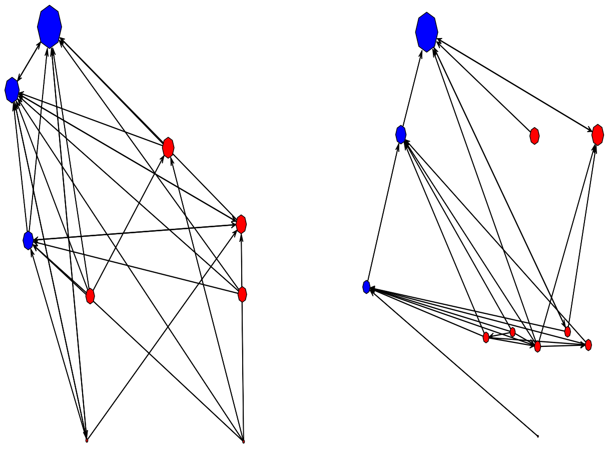

Fig. 11. Networks for high and low profit branches in a bipartite form. The left network corresponds to a low profit branch and the right one corresponds to a high profit branch. The blue nodes represent employees with high indegree and the red nodes represent employees with low indegree. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1. The variability in the Hay Points of the employees with high indegree is more in a high profit branch than in a low profit branch.
2. The variability in the Hay Points of employees going for advice to an employee with high indegree is more in a high profit branch than in a low profit branch.
3. The interaction among employees with low indegree across different levels of the branch is more in a high profit branch than in a low profit branch.

4.2. Quantifying the observed features of the network

To quantify these features, we use the notion of indegree centrality, one of the simplest measures used to determine the relative importance of a unit (in this case the employee represented by that node in the branch) in the graph; e.g., see Freeman [12]. The indegree centrality of a node is defined as the number of edges received by that node, i.e., the number of employees coming to that employee for advice. Indegree centrality for a group of nodes can be defined as the total number of edges received by that group from the other nodes in the graph not in the group.

Our aim is to measure the indegree centrality of the group of employees with high indegree, i.e., the group L. We also want to capture the variability in the number of Hay Points of the employees coming for advice with respect to the number of Hay Points of the employees in group L. We can do both of these things by using weighted indegree centrality as defined below:

$$C \equiv \sum_{i \in L} \sum_{j \in S} |H_i - H_j| e_{ji}$$

where H_i is the number of Hay Points for the ith employee and

$$e_{ji} = \begin{cases} 0 & \text{if no edge from node } j \text{ to node } i, \\ 1 & \text{otherwise.} \end{cases}$$

We can interpret C as the weighted indegree centrality of the employees in L where the weight for an edge from an employee in S to an employee in L is defined as the absolute difference of their Hay Points. The value of C increases as more and more employees in S from different levels of the branch come for advice to the employees in L. As we noted above, we also want to measure the interaction among employees with low indegree, i.e., group S across different levels of the branch. For this we again use

Table 4
Values of C and I for a low profit and a high profit branch, of the same size.

Branch	C	I
Low profit	799	175
High profit	1065	171

Table 5
Values of C and I for a low profit and a high profit branch of similar size.

Branch	C	I
Low profit	1063	224
High profit	1065	171

the total weighted indegree centrality of each of the employees in S restricted to the subgraph with nodes representing the employees in S, defined as follows:

$$I \equiv \sum_{i\in S}\sum_{j\in S} |H_i - H_j| e_{ji}.$$

We can interpret I as a measure of how much the employees in S have to move across different levels among themselves for advice. I measures the interaction within group S whereas C measures the interaction from S to L. In Table 4, we evaluate these measures for the two branches in Figs. 8 and 9.

Although the interaction in S is the same for the two branches, the interaction between S and L is much higher for the high profit branch. In Table 5, we evaluate the measures for the low profit and high profit branches in Figs. 10 and 11.

In this case, we can see that C is the same for both, but I is much higher in the low profit branch, i.e., the interaction in S in the low profit branch is much more than in the high profit branch. In Tables 4 and 5, we observe that profit increases with C and decreases with I. In order to compare the profit of all the branches we choose the following measure:

$$M \equiv \frac{I}{C}$$

where M is a measure for interaction in S per unit interaction from S to L. In other words, M measures how much the employees in S are moving among themselves for advice given that they are going to the employees in L for advice. According to our observation above, profit should decrease with M.

4.3. Justification of the choice of M for explaining profit

M is a measure for the underlying graph representing the network for the branch, but it does not take into account variability in profit due to size of the branch as well as the distribution of the number of Hay Points of the employees within each branch. As we have seen, there is considerable variability in the size as well as the Hay Point distribution for the branches. In order to study how profit varies with M across all the branches, we also have to account for the variability in profit due to size and Hay Point distribution. We do this through the sum of the number of Hay Points of all the employees in the branch, H, which accounts for both the size as well as the Hay Point distribution to a large extent.

In Fig. 12 we plot profit scaled by H versus M for all the branches, and we see that the scaled profit decreases linearly with M, although there are three visible outliers. Since we have already observed that I and C influence the profit of the branches, we might posit a model for predicting profit that includes I and C along with M. To examine the appropriateness of this model, we scale the profit of each branch by the total Hay Points of its employees, H, and fit two separate linear models. The first model has I, C and M as covariates and the second model has just M. Table 6 gives the values for AIC and BIC for the two models, and both criteria favor the simpler model which includes only M. This conclusion is supported by an examination of plots of the fitted scaled profit vs. the scaled profit for both models.

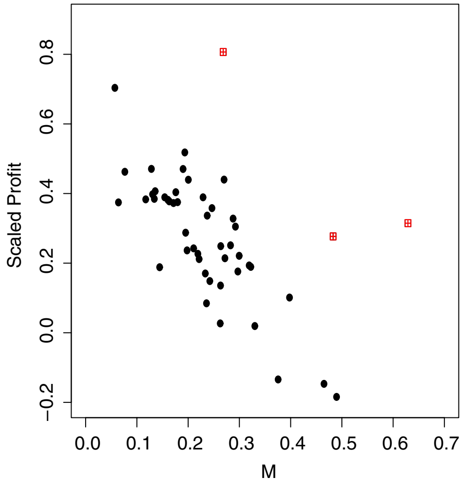

Fig. 12. Plot of scaled profit vs M for all the branches. There are three outliers, denoted by red squares.

Table 6
AIC and BIC for the two models.

Model predictors	AIC	BIC
I, C, and M	−85.10099	−75.95778
M	−72.13937	−66.65344

Table 7
Estimated coefficients with standard errors for the fitted model.

Coefficient	Estimate	Std. error	Conf. interval
$\hat{\alpha}$	1.21292	0.0464	[1.1544, 1.3067]
$\hat{\beta}_0$	0.1466	0.0500	[0.0710, 0.2290]
$\hat{\beta}_l$	−0.3641	0.1217	[−0.5650, −0.1794]

5. Model estimation and assessment of fit

Instead of predicting scaling profit as we did in the preceding section, here we choose to use the more general model

$$\text{Profit} = H^{\alpha}(\beta_0 + \beta_1 M) + \epsilon$$

where H is the sum of the Hay Points of all the employees in a branch and ϵ is the error with $E[\epsilon] = 0$ and $Var[\epsilon] = \sigma^2$.

5.1. Parameter estimation

The model proposed above is non-linear, and thus we use the Gauss–Newton algorithm for estimating the parameters by minimizing the mean squared error; cf. [27]. There are three outliers and we fit the proposed model excluding them.

We provide the estimates along with their standard errors and associated 95% confidence intervals in Table 7. We used a parametric bootstrap to estimate the standard errors since the residuals from the model do not seem to follow a normal distribution, as we can see in the QQ-plot of the residuals in Fig. 13. The distribution of the residuals has a heavier tail than a normal distribution.

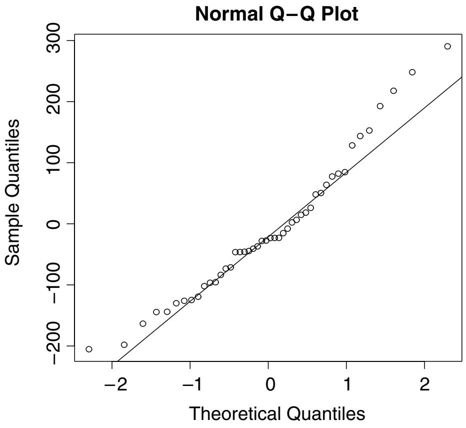

Fig. 13. *QQ*-plot for the residuals from the model.

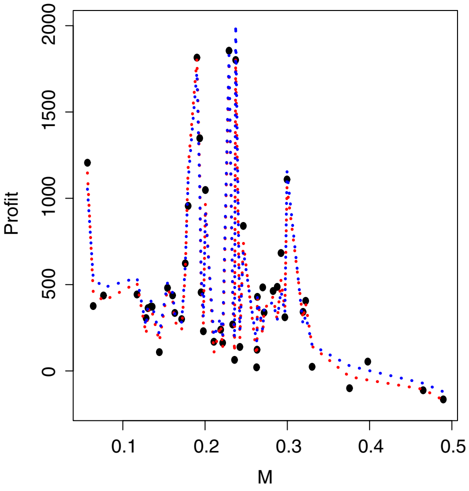

Fig. 14. Comparison of the fit from the nonparametric and the proposed model. The red line is for the nonparametric model and the blue line is for the proposed non-linear model.

5.2. Assessing goodness-of-fit

For testing the goodness-of-fit of the model, we use the nonparametric lack-of-fit test as proposed by Loader [20]. A nonparametric curve is fitted to the data with the total Hay Points and M as the two variables. We provide the details of this test in the Appendix. The p-value associated with the observed test statistics is 0.20 and we conclude that the model provides an acceptable fit to the data. We provide a comparison of the fitted curves for the proposed model and the nonparametric model in Fig. 14. We can see that the fits for the two models are quite close.

For predicting the profit of a branch given the data on its structural network for advice we thus use the following model:

$$\text{Profit} = H^{1.21}(0.1466 - 0.3641M). \tag{1}$$

The model (1) is non-linear, and we estimated the parameters in it using the Gauss–Newton algorithm to minimize the mean squared error.

6. Discussion

In this paper, we have explored the role of the advice network structure of organizations for prediction purposes. We worked with data gathered on employees in 52 branches of a major North American bank corporation, and we used the network structure discovered from exploratory analysis to predict the profitability of each of the branches.

Among the key insights emanating from our analyses are the following:

1. Employees doing different kinds of jobs need to communicate among themselves to perform efficiently.
2. If there are some employees with significantly different Hay Points who can give good advice then the other employees don't need to seek advice among themselves across different groups.
3. The statistic $M \equiv \frac{I}{C}$, which measures how much the employees in a group S are moving among themselves for advice given that they are going to the employees in L for advice, is small if the employees communicate well across different groups and the employees with high indegree are able to give good advice.
4. Profit (appropriately scaled) is a decreasing function of M.

Our ultimate model extracted a few small features and used them in a non-linear form. Early on in our analyses we explored the use of exponential random graph models (ERGMs) for the network for each branch using tools in the *statnet* package developed at the University of Washington [15]. This provided many of the insights into how the employees worked which led to the prediction models that we ultimately examined. Nonetheless, many of these ERGMs displayed the degeneracies and near degeneracies described by Handcock [14] and Rinaldo et al. [22]. In the near future we hope to return to these more formal network model structures and to study the dependence of bank branch profit on the estimated ERGM distributions for the advice network within each branch.

Appendix. Nonparametric lack-of-fit test

Loader [20, Sec. 4.3] describes a lack-of-fit test that can be used without repeated observations or prior knowledge of σ^2 based on comparing the fit of the parametric model to the fit of a smoother. For each data point, we find the fitted value $\hat{y}_i$ from the parametric fit and $\tilde{y}_i$, the fitted value from the smoother. If the parametric model is appropriate for the data, then the differences $(\hat{y}_i - \tilde{y}_i)$ should all be relatively small. A suggested test statistic is based on looking at the squared differences and then dividing by an estimate of σ^2,

$$G = \frac{\sum_{i=1}^{n}(\hat{y}_i - \tilde{y}_i)^2}{\hat{\sigma}^2}$$

where $\hat{\sigma}^2$ is the estimate of variance from the parametric fit. Large values of G provide evidence against the NH that the parametric mean function matches the data. Loader [20] provides a bootstrap for computing an approximate significance level for a test based on G.

The appropriate bootstrap algorithm is a little different from what we have seen before and uses a parametric bootstrap. It works as follows:

1. Let $y_i,\ i = 1, \ldots, n$, be the response used for regression.
2. Fit the parametric and non-parametric model to the data, and compute G for each model. Save the residuals, $\hat{e}_i = y_i - \hat{y}_i$ from the parametric fit.
3. Obtain a bootstrap sample $\hat{e}_1^*, \ldots, \hat{e}_n^*$ by sampling with replacement from $\hat{e}_1, \ldots, \hat{e}_n$. Some residuals will appear in the sample many times, some not at all.
4. Given the bootstrap residuals, compute a bootstrap response Y^* with elements $y_i^* = \hat{y}_i + \hat{e}_i^*$. Use the original predictors unchanged in every bootstrap sample. Obtain the parametric and nonparametric fitted values with the response Y^*, and then compute G.
5. Repeat steps 2 to 3 B times.
6. The significance level of the test is estimated to be the fraction of bootstrap samples that give a value that exceeds the observed G.

References

[1] E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing, Mixed membership stochastic blockmodels, Journal of Machine Learning Research 9 (2008) 1823–1856.

[2] R. Albert, A. Barabási, Statistical mechanics of social networks, Review of Modern Physics 74 (2002).

[3] A.-L. Barabási, Linked: The New Science of Networks, Perseus, Cambridge, MA, 2002.

[4] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.

[5] M. Buchanan, Nexus: Small Worlds and The Groundbreaking Science of Networks, W.W. Norton, New York, 2002.

[6] N.A. Christakis, J.H. Fowler, Connected: The Surprising Power of our Social Networks and How they Shape our Lives, Little, Brown, and Co., 2009.

[7] R. Cross, A. Parker, The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations, Harvard Business School Press, 2004.

[8] P.S. Dodds, R. Muhamad, D.J. Watts, An experimental study of search in global social networks, Science 301 (2003) 827–829.

[9] R. Durrett, Random Graph Dynamics, Cambridge University Press, New York, 2007.

[10] S.E. Fienberg, M.M. Meyer, S.S. Wasserman, Statistical analysis of multiple sociometric relations, Journal of the American Statistical Association 80 (1985) 51–67.

[11] O. Frank, D. Strauss, Markov graphs, Journal of the American Statistical Association 81 (1986) 832–842.

[12] L.C. Freeman, Centrality in social networks: Conceptual clarification, Social Networks 1 (3) (1978) 215–239.

[13] A. Goldenberg, A. Zhang, S.E. Fienberg, E. Airoldi, A survey of statistical network models, Foundations and Trends in Machine Learning 2 (2) (2010) (in press).

[14] M.S. Handcock, Assessing degeneracy in statistical models of social networks, Working paper No. 39, Center for Statistics and the Social Sciences, University of Washington, 2003.

[15] M.S. Handcock, D.R. Hunter, C.T. Butts, S.M. Goodreau, M. Morris, Statnet: Software tools for the statistical modeling of network data, Seattle, WA, Version 2.0, 2003, URL: http://statnetproject.org.

[16] M.S. Handcock, A.E. Raftery, J. Tantrum, Model-based clustering for social networks (with discussion), Journal of the Royal Statistical Society, Series A 170 (2007) 301–354.

[17] P.W. Holland, S. Leinhardt, An exponential family of probability distributions for directed graphs (with discussion), Journal of the American Statistical Association 76 (1981) 33–65.

[18] D. Hunter, M.S. Handcock, Inference in curved exponential family models for networks, Journal of Computational and Graphical Statistics 15 (2006) 565–583.

[19] D. Krackhardt, D. Brass, Intra-organizational networks: The micro side, in: S. Wasserman, J. Galaskiewicz (Eds.), Advances in the Social and Behavioral Sciences from Social Network Analysis, Sage, Beverley Hills, CA, 1994, pp. 209–230.

[20] C. Loader, Smoothing: Local Regression Techniques, in: Handbook of Computational Statistics: Concepts and Methods, Springer-Verlag, Heidelberg, 2004, pp. 539–562.

[21] M.E.J. Newman, S.H. Strogatz, D.J. Watts, Random graphs with arbitrary degree distributions and their applications, Physics Review E 6402 (026118) (2001).

[22] A. Rinaldo, S.E. Fienberg, Y. Zhou, Maximum likelihood estimation in discrete exponential families with application to exponential random graph models, Electronic Journal of Statistics 3 (2009) 446–484.

[23] D. Strauss, M. Ikeda, Pseudolikelihood estimation for social networks, Journal of the American Statistical Association 85 (1990) 204–212.

[24] H. Tosi, Theories of Organization, Sage, Thousand Oaks, CA, 2008.

[25] D.J. Watts, Six Degrees: The Science of A Connected Age, J.J. Norton, New York, 2003.

[26] D.J. Watts, Small Worlds: The Dynamics of Networks Between Order and Randomness, Princeton University Press, Princeton, NJ, 1999.

[27] S. Weisberg, Applied Linear Regression, 3rd ed., Wiley, New York, 2005.