Palestine Journal of Mathematics

Vol 13(Special Issue III)(2024) , 136–147

SOCIAL NETWORK DATA ANALYSIS USING SEMIPARAMETRIC REGRESSION

Sreeja V. N., P G Sankaran, Deepa V. G. and Maya S. S.

Communicated by Kuncham S.P

MSC 2010 Classifications: Primary 91D30; Secondary 62Nxx.

Keywords and phrases: Networks, Social networks, Survival analysis, Semiparametric regression.

Abstract The fastest-growing fields of research in the modern world include social networks and their associated studies. This field is famed for being a highly effective way to investigate interpersonal connections. We tried to implement a creative strategy to improve institutional administrative activities. If we need to introduce a new concept within our organization, we have many alternatives, one of which is a "train the trainer" program. In an attempt to train almost the whole staff of an institution about software, we tried to understand the interactions between the staff members as a network. We use semi-parametric regression analysis to identify the covariates affecting each person's training. It enables us to examine how various factors impact the probability of receiving training at a specific time.

1 Introduction

The idea of social networks has garnered considerable interest from the behavioral and social science communities in recent years. Much of this interest can be attributed to the engaging study of the relationships among various social entities. The concept of social networks allows researchers to study the various social relationships that exist within a society. Through social network analysis (SNA), one can identify various social concerns and develop effective strategies to address social relationships. It is a fast-expanding multidisciplinary field that includes social, demographical, mathematical, statistical, medical, and computer sciences. The term "social networks" refers to the social relationships that exist between individuals, families, groups, and communities. Investigators in areas such as sociology, economics, medicine, psychology, etc. are keen to examine a society by examining various social concerns and the pattern of relationships within it. The researcher can use SNA to show a variety of linkages between individual groups, such as information flow among employees.A book by [1] can assist researchers in learning about the many different social network methods available, comprehending the theoretical foundations of these techniques, and finding some direction in terms of choosing the methods that are most suitable for a particular set of challenges. Social network analysis is ideal for assessment methods in social studies, as discussed in the chapter by [2]. The authors of [3] evaluated the topics that social researchers have attempted to explain using social network analysis and gave a concise summary of the fundamental presuppositions, objectives, and explanatory mechanisms that are common in the discipline. A leading basic book on the topic of networks was authored by [4]. It offers a strong foundation and appeals to people from many different perspectives and academic fields. Social networks with thousands of nodes and links may now be extracted and analyzed. Instead of employing conventional social network research techniques, the book [5] strives to present innovative ways to investigate complicated data sets. In [6], the authors attempted to examine by representing data from a tribal village's interaction throughout times of normalcy and crisis as a directed graph. The overview of social network analysis in [7] begins with an introduction and then discusses several ways to describe social networks, including graphs, formal approaches, and matrices. They introduced several tools used for the study and visualization of social networks to examine and mine the data from such networks.

The study of occurrences of events that include heterogeneous data and time elements are significant issue with numerous applications in a variety of fields. The phrase "survival data"

refers to information that measures the time that passes before an event takes place. A collection of techniques for analyzing such data is called survival analysis, often known as event history analysis or duration analysis. The event denotes a change from one status to another. In survival analysis, the event is generally death, although time to recover from any illness is also an event. Events in economics and demography include marriage and receiving the very first recruitment notification for a dream field. The method of [8] covers the nature of the problem that survival analysis attempts to solve, the outcome variable taken into account, the necessity of accounting for censored data, what a survival function and a hazard function represent, fundamental data layouts for survival analysis, the objectives of survival analysis, and some examples of survival analysis. The strategy used in [9] has the intention of offering a focused text on regression modeling for the time-to-event data frequently found in studies related to health. All of the previously covered topics include univariate and independent analysis that is time-dependent. By allowing for multivariate times, [10] expands the field, and in-depth descriptions are provided for the concepts and possible future data types. The analysis of such data using various methods is presented from an applied perspective. In addition to these sorts of studies, several statistical approaches are often employed by researchers in diffusion studies to assess various contributing elements. [11] presented a framework for analyzing the features of the social network diffusion process by incorporating competing risks into survival analysis. One of the main assumptions in this sort of research is the homogeneity of the study population, which is not always achievable in real-world circumstances. Individuals differ greatly from one another in reality. There may be specific factors (covariates) that impact an event's occurrence. The survival times in standard survival models are assumed to be independent of one another. In the book [12], the author attempts to describe the applications of survival analysis using R software. The book [13] has a strong emphasis on problem-solving techniques that handle the numerous problems that might occur while creating multivariate models using actual data. The reader will have a thorough comprehension of prediction accuracy as well as the negative effects of classifying ongoing predictors or results. Additionally, it offers a variety of graphical techniques for explaining complicated regression modes.

2 Objective of the study

There are several approaches we may take if we need to introduce a new subject or program in an organization. One of them is the "train the trainer" program, which is needed to introduce a new subject in an organization. Using an outside source to train an organization's entire workforce is both impossible and expensive. Furthermore, independent specialists may be unable to attend to the occasional review and resolution of issues. In such circumstances, this method comes in handy. People may be competent in their field but have less ability to train others in the new area. What is being done in this training program is to provide knowledge about the topic which the organizers want to share among all workers and teaching abilities to efficient individuals who must transmit information to other coworkers in a repeatable manner. The knowledge, performance, and efficiency of the organization's members may all be improved by the dissemination of information in this way. The first phase of this training program involves training an individual or a group.

It may take hours, days, or even months to complete this training. After receiving training, attendees can impart their knowledge to other staff members within their organization. This individual again trains a few more people in a diffusional fashion. Until all employees in the organization have undergone training, you may continue this practice. Multilevel trainers may be present in this training program. Through this kind of training program, an effort is made to provide training to an institution's entire staff. Some people made an effort to share their knowledge with others as soon as they could. However, some people took longer to diffuse the information because of various factors. We attempt to understand the effect of these factors through semi-parametric regression analysis. Using social network analysis, it seeks to identify individuals who have the greatest influence over fellow employees.

3 Terminology

In social networks, Individuals, families, and groups are referred to as nodes, actors, or vertices whereas edges or linkages between the nodes refer to the connections between people. Graphs can be used to represent the connection for analytical reasons. If V denotes the vertices set and E denotes the edges set (directed) then $G = (V, E)$ will be a directed graph. Graphs are very useful for visualizing a wide range of social circumstances. Theoretically, graphs and networks are equivalent, although the term "network" is more commonly used when discussing social or technical relationships. Individual nodes' social capital can be measured using social network diagrams. A social network's structure helps in understanding the relationships among its members.

Let us refer to the individual who completed training as an adopter and a receptive person who is eligible to get coaching from his or her coworker quite soon. If a receptive individual received training from a past adopter, a receptive edge exists between them. If the individuals in the institution are viewed as vertices, the relationship (got knowledge after training) among them and the individuals in the institution jointly create the directed graph $G = (V, E)$ with V as the set of adopters and E is the set of directed edges; $(i, j) \in E$ for $I, j \in V$. This means that personal j was trained as a result of the individual i. A considered event is a receptive person who received training from a prior adopter.

4 Methodology

Our present study is regarding such a good training program at an educational institution in Kerala State. This institution has 70 faculty members and around 1600 students. They intend to deploy new software to manage students' personally identifiable information, attendance, and other academic administration-related data. They thought about providing excellent training to those faculty members at a reasonable cost. If a software development trainer provided instruction to all faculty members at the same time, experts in this field may be unable to attend to the occasional evaluation and settlement of challenges. In such cases, the "train the trainer" program is beneficial. What is done here is to give training skills to efficient employees who must transmit information to other coworkers in a repeatable manner. A principal trainer is someone who received their initial training from outside experts. Secondary trainers absorb information from the primary trainer; they then teach secondary-level (here, we take four faculty members as secondary-level) faculty members how to teach others. Third-level trainers include knowledge from secondary-level trainers, and so on until all members of the institution have been trained. This procedure was continued until all 70 faculty members were trained. Figure 1 indicates the sociogram connecting coworkers of the Institution.

4.1 Social network method

A subset D of V is referred to as a dominating set in graph theory for a graph $G = (V, E)$ if every vertex that is not in D is adjacent to at least one member of D. The dominance number of a graph G is the total number of vertices that constitute its minimal dominating set. It is indicated by $\gamma(G)$. A directed graph is made up of a finite number of elements $(V_1, V_2, ...V_n)$ and a finite number of ordered pairs (V_i, V_j) of those elements, with no repeated ordered pair. The set's components are referred to as the directed graph's vertices, and the ordered pairings are referred to as its directed edges. We utilize notation to show that the edge (V_i, V_j) is a member of the directed graph. For each different pair of vertices, V_i and V_j, either $V_i \to V_j$ or $V_j \to V_i$ holds, but not both. This is known as a dominance-directed graph. Tournaments is another term for this sort of graph. Assume that the relationship between V_i and V_j represents the idea that "V_i influence the person V_j ". Finding the most influential point is then the challenge. It may be demonstrated that there is at least one vertex in a dominance-directed graph from which there is a one- or two-step link to any other vertex. Associate the nxn matrix M with the graph G if it contains n vertices.

$$\mathbf{m}_{ij} = \begin{cases} 1 & \text{if } V_i \longrightarrow V_j \ , \\ 0 & otherwise \end{cases} \tag{4.1}$$

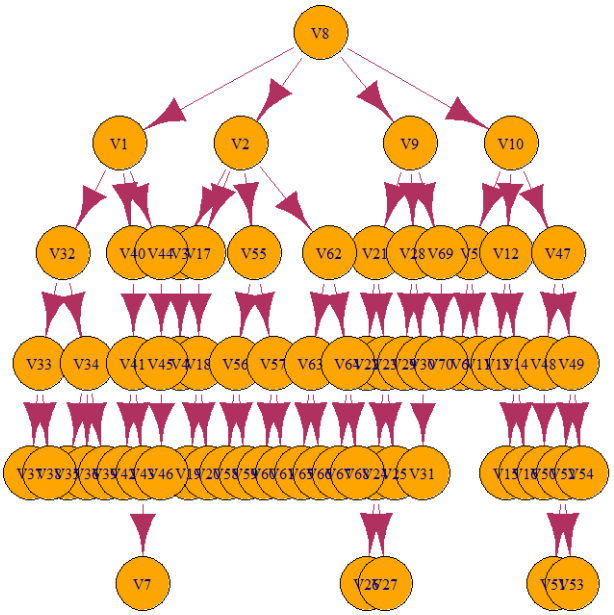

Figure 1. Sociogram representing the training pattern among coworkers of the institution

All elements in the vertex matrix seem to be either zero or one, and all diagonal elements are zero. The matrix uniquely determines the connected network. Therefore, to find the vertices with the greatest influence, we must compute $M + M^2$. To demonstrate the power of the associated vertex, add the values of each row of $M + M^2$; the biggest number indicates the vertices with the most influence.

4.2 Regression method

The survival function, which represents the probability that the person has still not experienced the event until time t, is the fundamental quantity being used to characterize time-to-event events. It is given by

$$S(t) = P(T > t) = 1 - F(t) \tag{4.2}$$

The hazard function serves as the foundation for survival analysis. It represents a person's imminent risk of experiencing the event of interest if the person has never observed the event. The hazard function is defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T > t)}{dt} \tag{4.3}$$

The cumulative hazard function is

$$\Lambda(t) = \int_0^t \lambda(t)dt = -ln[S(t)] \tag{4.4}$$

$\Lambda(t)$ uniquely determines the distribution uniquely by

$$S(t) = exp[-\Lambda(t)] \tag{4.5}$$

Data in survival studies may be incomplete for a variety of reasons. Censorship is one of the causes of incompleteness. Censored data frequently appears in survival studies since the investigator is unable to gather detailed information on each event's occurrence times. The experiment may have ended before all participants witnessed the event, or the participant may have been

dropped from the study at some point. The existence of censored observations hampers time-dependent data analysis. Right, left, type I, type II, and other forms of censorship are a few kinds of censorship. A Type I censoring procedure happens if each participant has a predetermined censoring time $C_i > 0$ so that when the event happens T is viewed only if $T \leq C_i$ otherwise, we only know that $T > C_i$. When a study is conducted over a defined time period, Type I censoring is typical.

If $T_1, T_2, ...T_n$ are independent and identically distributed random variables with common probability density function $f(t)$ and survival function $S(t)$, then the likelihood function under Type I censoring is

$$L = f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{4.6}$$

where $\delta_i = I(T_i = t_i) = \begin{cases} 1 & \text{if } T_i = t_i \\ 0 & \text{if } T_i > t_i \end{cases}$, for $i = 1, 2, ...n$ is known as status or censoring indicator of t_i from which we can understand is an observed time or censored time.

One of the goals of survival research is to estimate the survival function. There are numerous ways to do it, including parametric, nonparametric, and semiparametric methods. In the parametric method, the survival time seems to have a probability function with an unknown parameter, while in the nonparametric approach, the survival function is analyzed using the Kaplan-Meier estimate. The estimator provided by Nelson-Aalen could be utilized to directly estimate the cumulative hazard function. Regression modeling is used to study the relationship between covariates and event occurrence times. The inclusion of variables in a regression model is an essential mechanism for describing population heterogeneity. Recognizing the connection between variables as well as event occurrence time is the main objective of these studies. The proportional hazards model proposed by Cox is described in [14], which evaluates the influence of covariates mostly on the hazard function and is a popular regression model in survival studies. According to the proportional hazard principle, the hazard ratio between any two people stays unchanged throughout time. This is an excellent choice whenever a model is adequately stated and all key factors are taken into consideration. In reality, all important factors can rarely be accounted for. Explanatory factors are then used to account for both apparent and hidden variability. The proportional hazards model is

$$\lambda(t|\bar{z}) = \lambda_0(t) e^{\bar{\beta}' \bar{z}}, \quad t > 0 \tag{4.7}$$

where $\lambda_0(t)$ is baseline hazard function and $\bar{\beta}$ is the parameter vector. As it includes the unspecified baseline hazard function as well as the parametric vector $\bar{\beta}$, the model is semi-parametric. The partial likelihood function for estimating the parametric vector $\bar{\beta}$ is

$$L(\bar{\beta}) = \prod_{i=1}^{n} [\frac{e^{\bar{\beta}' \bar{z}}}{\sum_{j=1}^{n} Y_j(t_i) e^{\bar{\beta}' \bar{z}}}]^{\delta_i} \tag{4.8}$$

where $Y_i(t) = I(t_i \geq t), i = 1, 2, ...n$ is an indicator function. The partial likelihood is then maximized to yield an estimate of $\bar{\beta}$. A nonparametric estimator for the survival function using censored data was developed by Kaplan and Meier. The baseline hazard function may be estimated using the generalized Nelson-Aalen estimator. The cumulative baseline hazard function's maximum likelihood estimate was provided by Breslow. This estimator is commonly applied in many applications. The estimator is very helpful in the development of Cox regression and semiparametric inference with censored data.

5 Results and discussion

To better understand the data, we analyse it from two different perspectives: social network analysis and the impact of covariates in training-induced knowledge acquisition.

Table 1. Row total of the vertex (node) matrix $M + M^2$

Row number	1	2	3	4	5	6	7	8	9	10
Row total	6	8	2	0	4	0	0	8	6	6
Row number	11	12	13	14	15	16	17	18	19	20
Row total	0	4	0	4	0	0	2	4	0	0
Row number	21	22	23	24	25	26	27	28	29	30
Row total	4	0	4	4	0	0	0	4	0	2
Row number	31	32	33	34	35	36	37	38	39	40
Row total	0	4	4	6	0	0	0	0	0	2
Row number	41	42	43	44	45	46	47	48	49	50
Row total	4	0	2	2	2	0	4	2	4	0
Row number	51	52	53	54	55	56	57	58	59	60
Row total	0	4	0	0	4	4	4	0	0	0
Row number	61	62	63	64	65	66	67	68	69	70
Row total	0	4	4	4	0	0	0	0	2	0

5.1 Social network method

On the occasion of implementing any software, everyone will get the necessary time to learn it. However, once you begin using that one, you will need to upgrade the program to meet the demands of the situation in many circumstances. In such cases, the software upgrade modifications must be taught to all organization staff members in a relatively short period. One method is to inform all employees of the software upgrade changes on one day and provide them with the appropriate training. However, organizing a program like this for everyone in a running institution may be unfeasible. Here comes the significance of the dominating set. If you can identify a minimum dominant set, the ideal one-step strategy is to train the dominant set members first and then train the others through them. The shortest dominant set D, which is a subset of the vertex set V for graph $G = (V, E)$ for our data, is as follows:

$D = (V_1, V_2, V_3, V_5, V_7, V_8, V_9, V_{10}, V_{12}, V_{14}, V_{18}, V_{21}, V_{23}, V_{24}, V4_{28}, V_{30}, V_{33}, V_{34}, V_{41}, V_{43}, V_{45}, V_{47}, V_{48}, V_{49}, V_{52}, V_{55}, V_{56}, V_{57}, V_{62}, V_{63}, V_{64}, V_{69})$

The dominance number $\gamma(G) = 32$ .Thus, by utilizing the dominant set in this "train the trainer program" we may manage a delicate issue as if it were simple. The benefit of this is that there won't be any disturbance to the organization's regular working when the software upgrades are being sent to everyone who seeks the help of these members of the dominating set. The power of $V_1 = 6, V_2 = 8$, and so on can be obtained from Table 1. It is apparent that V_2 and V_8 are the most powerful vertices, or that the second and eighth individuals have more influence on the institution's other faculty members. So, if we need to inform staff people about some development of the software quickly, we may employ the second and eighth persons.

5.2 Regression method

The heterogeneous and time-dependent data that we are discussing here may be analyzed using survival analysis. The time during which people receive training to learn about the software is referred to as "event time." The goal of survival analysis is to analyze the influence of factors on survival time. In other words, it allows us to look at how various factors influence as quickly a specific event happens at a certain time. Predictor variables are often referred to as covariates (or factors). The stream (Science, Arts, and Commerce), service time (year of experience), gender, and age of faculty members are considered covariates in our study and are elements that may affect training. The institution's authority decided to provide this software training to all staff members within 30 days. Some people did not receive training on time. They should have

Table 2. Inference by using Cox proportional hazards model

Cox proportional hazards					
	coef	exp(coef)	se(coef)	z	p
Gender	0.4164	1.5164	0.2823	1.475	0.14
Likelihood ratio test=2.29 on 1 df, p=0.1306					
	coef	exp(coef)	se(coef)	z	p
Service period	0.08896	1.09304	0.03515	2.531	0.0114
Likelihood ratio test=6.24 on 1 df, p=0.01247					
	coef	exp(coef)	se(coef)	z	p
Stream	-0.2899	0.7484	0.1308	-2.216	0.0267
Likelihood ratio test=4.84 on 1 df, p=0.02777					
	coef	exp(coef)	se(coef)	z	p
Age	0.01901	0.01919	0.02064	0.0921	0.357
Likelihood ratio test=0.85 on 1 df, p=0.3565					

been trained by individuals who had just undergone training. The delay in receiving training might be due to a variety of factors affecting the people who should be taught or receive the training. As training must be completed within 30 days, every day obtained beyond that is considered censoring time. Our data was then analysed using R programming and the Kaplan-Meier estimator.

We obtained Table 2 by using the "coxph" function in R programming and the Breslow method. Gender and age do not affect training because the p-value is large ($p > 0.05$), but service period and stream do. In other words, gender and age do not play a role in the delay in getting training for each faculty member. This training time, however, has an impact on the service period of each faculty and stream to some extent.

In our study, even though there were 70 faculty members, 65 of them learned how to use the software after receiving training within 30 days. Since the study period ends in 30 days, we are unable to monitor the occurrence of the other 5 individuals' events, and hence they were considered censored. The survival probabilities for each person with a varied service period are explained in Table 3. At the end of four days, 16 people are at risk, and one occurrence has occurred. The likelihood of surviving for more than two days is 93.75%, with a standard error of 0.0605. That is, the probability of being taught about the software in more than two days is 93.75%. Alternatively, the 93.75% confidence interval for survival is (82.609%, 100%). The same interpretation applies to other values.

In Figure 2, SP1 denotes those with a 1-year service experience, SP2 denotes those with a 2-year service experience, and so on. Individuals with service periods of eight received their training earlier on average. Similar to this, most of them with service period 1 receive training only after 20 days.

In Table 4, after two days, 31 individuals are still at risk, and one event has already happened. 96.77% of people are likely to get training for more than two days, with a standard error of 0.0317. We may similarly comprehend other values.

From Figure 3, it is clear that faculties in the commerce stream got trained sooner as compared to arts and science faculties.

Figure 4 demonstrates how male faculty members received training more quickly than female faculty.

Table 3. Service period-wise calculation of survival time

Service period=1 year						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
4	16	1	0.9375	0.0605	0.82609	1
9	15	3	0.75	0.1083	0.5652	0.995
11	12	2	0.625	0.121	0.42761	0.914
12	10	1	0.5625	0.124	0.36513	0.867
15	9	1	0.5	0.125	0.30632	0.816
17	8	1	0.4375	0.124	0.25101	0.763
18	7	1	0.375	0.121	0.19921	0.706
22	6	1	0.3125	0.1159	0.15108	0.646
25	5	1	0.25	0.1083	0.10699	0.584
26	4	1	0.1875	0.0976	0.06761	0.52
28	3	1	0.125	0.0827	0.03419	0.457
32	2	1	0.0625	0.0605	0.00937	0.417
Service period=2 years						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
9	4	1	0.75	0.217	0.4259	1
16	3	1	0.5	0.25	0.1877	1
17	2	1	0.25	0.217	0.0458	1
Service period=3 years						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
12	8	1	0.875	0.117	0.6734	1
15	7	1	0.75	0.153	0.5027	1
16	6	1	0.625	0.171	0.3654	1
17	5	1	0.5	0.177	0.25	1
18	4	1	0.375	0.171	0.1533	0.917
19	3	1	0.25	0.153	0.0753	0.83
22	2	1	0.125	0.117	0.02	0.782
Service period=4 years						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	3	1	0.667	0.272	0.2995	1
16	2	1	0.333	0.272	0.0673	1
17	1	1	0	NaN	NA	NA
Service period=5 years						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
8	4	1	0.75	0.217	0.4259	1
9	3	1	0.5	0.25	0.1877	1
10	2	1	0.25	0.217	0.0458	1
15	1	1	0	NaN	NA	NA
Service period=7 years						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	12	1	0.917	0.0798	0.773	1
3	11	2	0.75	0.125	0.541	1
8	9	2	0.583	0.1423	0.362	0.941
10	7	2	0.417	0.1423	0.213	0.814
11	5	1	0.333	0.1361	0.15	0.742
12	4	2	0.167	0.1076	0.047	0.591
16	2	2	0	NaN	NA	NA
Service period=8 years						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	9	1	0.889	0.105	0.7056	1
5	8	2	0.667	0.157	0.42	1
8	6	3	0.333	0.157	0.1323	0.84
9	3	1	0.222	0.139	0.0655	0.754
10	2	2	0	NaN	NA	NA
Service period=10 years						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	14	1	0.929	0.0688	0.803	1
4	13	2	0.786	0.1097	0.5977	1
5	11	3	0.571	0.1323	0.363	0.899
9	8	2	0.429	0.1323	0.2341	0.785
11	6	1	0.357	0.1281	0.1769	0.721
12	5	1	0.286	0.1207	0.1248	0.654
15	4	1	0.214	0.1097	0.0786	0.584
29	3	1	0.143	0.0935	0.0396	0.515

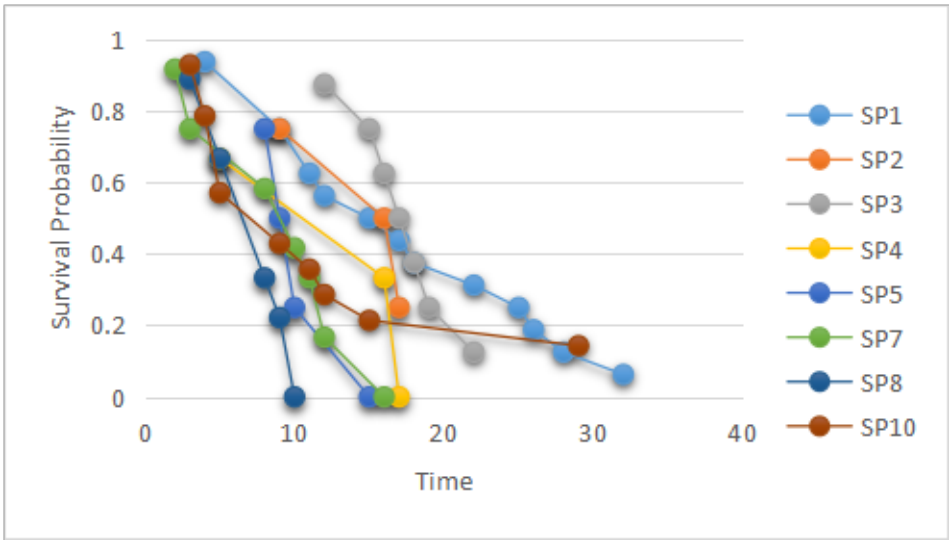

Figure 2. Service period-wise comparison of survival probability

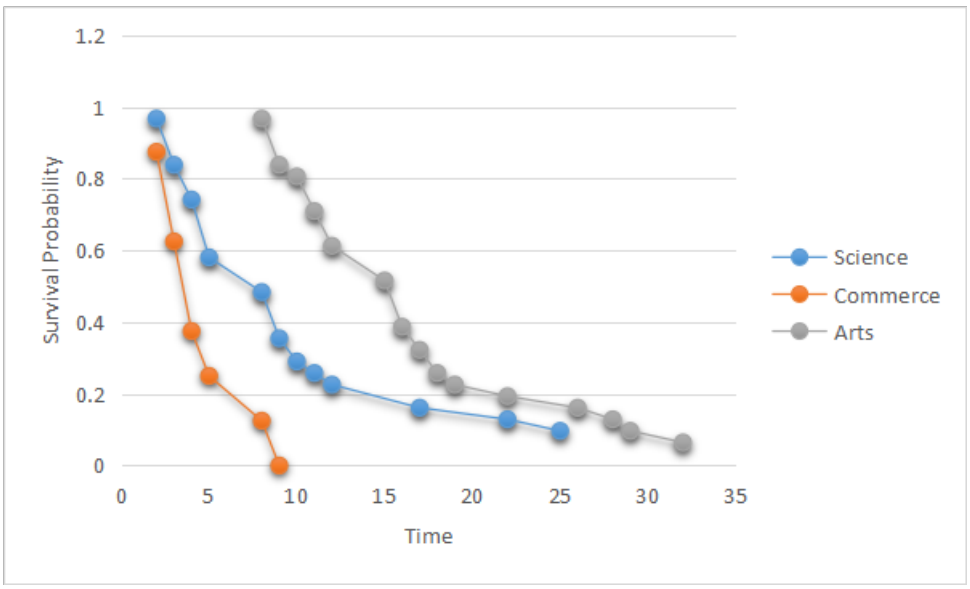

Figure 3. Stream-wise comparison of survival probability

Table 4. Stream-wise calculation of survival time

Science						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	31	1	0.9677	0.0317	0.9075	1
3	30	4	0.8387	0.0661	0.7187	0.979
4	26	3	0.7419	0.0786	0.6028	0.913
5	23	5	0.5806	0.0886	0.4305	0.783
8	18	3	0.4839	0.0898	0.3364	0.696
9	15	4	0.3548	0.0859	0.2207	0.57
10	11	2	0.2903	0.0815	0.1674	0.503
11	9	1	0.2581	0.0786	0.1421	0.469
12	8	1	0.2258	0.0751	0.1177	0.433
17	7	2	0.1613	0.0661	0.0723	0.36
22	5	1	0.129	0.0602	0.0517	0.322
25	4	1	0.0968	0.0531	0.033	0.284
Commerce						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	8	1	0.875	0.117	0.6734	1
8	7	2	0.625	0.171	0.3654	1
10	5	2	0.375	0.171	0.1533	0.917
12	3	1	0.25	0.153	0.0753	0.83
15	2	1	0.125	0.117	0.02	0.782
16	1	1	0	NaN	NA	NA
Arts						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
8	31	1	0.9677	0.0317	0.9075	1
9	30	4	0.8387	0.0661	0.7187	0.979
10	26	1	0.8065	0.071	0.6787	0.958
11	25	3	0.7097	0.0815	0.5666	0.889
12	22	3	0.6129	0.0875	0.4633	0.811
15	19	3	0.5161	0.0898	0.3671	0.726
16	16	4	0.3871	0.0875	0.2486	0.603
17	12	2	0.3226	0.084	0.1937	0.537
18	10	2	0.2581	0.0786	0.1421	0.469
19	8	1	0.2258	0.0751	0.1177	0.433
22	7	1	0.1935	0.071	0.0943	0.397
26	6	1	0.1613	0.0661	0.0723	0.36
28	5	1	0.129	0.0602	0.0517	0.322
29	4	1	0.0968	0.0531	0.033	0.284
32	3	1	0.0645	0.0441	0.0169	0.247

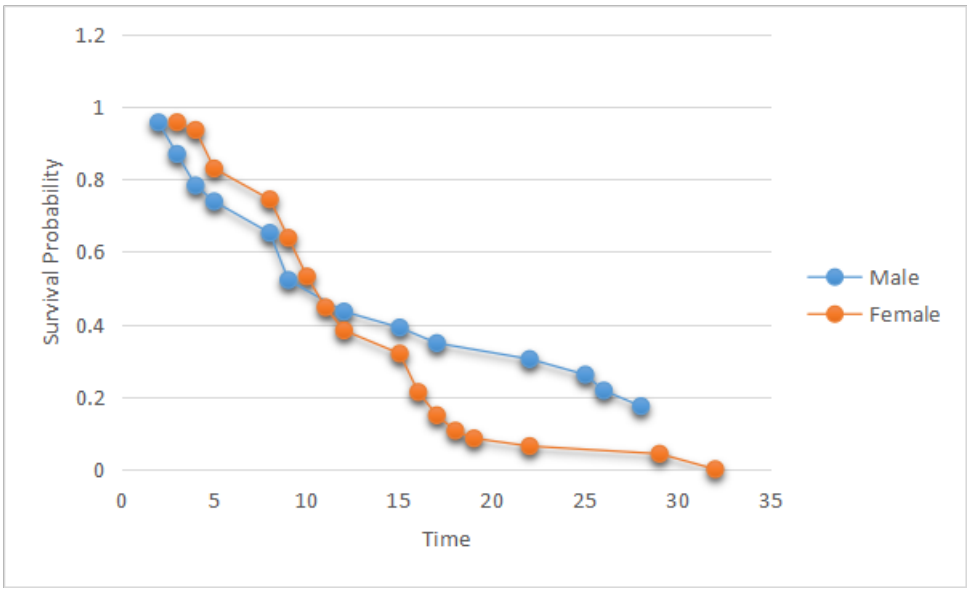

Figure 4. Gender-wise comparison of survival probability

6 Conclusion

We attempted to demonstrate how to effectively teach something new to workers in any organization at a low cost by using the "Train the Trainer" program. Its potential was investigated using social network analysis. Each employee was treated as a node in the study, and "training" was treated as an edge. Understanding the dominant set revealed that information can be delivered to everyone as soon as possible with their assistance when further training is required at some point. Similarly, it was discovered that node 2 and node 8, which are the most influential individuals, can be relied on to provide information quickly. The authority's initiative to train all staff members in one month was not carried out, at least in some cases. Its causes were investigated using semiparametric regression analysis. It was discovered that gender and age differences are not reasons to postpone training. However, each employee's service period and stream have an impact.

References

[1] Wasserman S., and Faust K, Social Network Analysis: Methods and Applications, Cambridge University Press, New York, (1994).

[2] Maryann Durland and Kimberly A. Fredericks, An Introduction to Social Network Analysis, New Directions for Evaluation (107):5 - 13, (2005).

[3] Borgatti, Mehra, Brass, and Labianca, Network analysis in the social sciences,*Science* Volume **323,** No.5916 (2009), 892-895.

[4] Mark Newman, Networks, Second Edition, Oxford University Press, (2018).

[5] Katarzyna Musial, Piotr Brodka and Matteo Magnani, Social Network Analysis in Applications, *Ai communications* Volume **29,** issue 1, (2015), 55-56.

[6] Deepa V. G., Aparna Lakshmanan S., and Sreeja V. N., Centrality and reciprocity in directed social networks: A Case study, *Malaya Journal of Matematik* Volume **S,** issue 1, (2019), 479-484.

[7] Jyoti Shokeen, Partibha Yadav, Overview of Social Network Analysis and Tools, *Journal of emerging technologies and innovative research* Volume **3,** issue 8, (2016), 29-31.

[8] David G. Kleinbaum, Mitchel Klein Survival Analysis: A Self-Learning Text, Springer Verlag, New York,3rd Edn, (2012).

[9] David Hosmer, Stanley Lemeshow, Susanne May Applied Survival Analysis: Regression modeling of time to event data, Wiley,2nd Edn, (2008).

[10] Jiacheng Wu, Forrest W Crawford, David A Kim, Derek Stafford and Nicholas A Christakis Exposure, hazard and survival analysis of diffusion on social networks,*Statistics in Medicine* Volume **37,** issue 8, (2018), 2561-2585.

[11] Hauguard P Analysis of Multivariate Survival Data, Springer-Verlag, New York, (2000).

[12] Dirk. F. Moore Applied survival analysis using R, Springer International Publishing (2016).

[13] Frank, E Harrell, Jr. Regression modeling strategies: With applications to linear models, Logistic and ordinal regression and survival analysis, Springer International Publishing, (2015).

[14] Cox, D. R. Regression models and life tables,*Journal of the Royal Statistical Society* Volume **34,** issue B, (1972), 187-220.

Author information

Sreeja V. N., Department of Mathematics, Sree Krishna College, Guruvayur - 680104, Kerala., India.
E-mail: drsreejavn@gamil.com

P G Sankaran, Department of Statistics, Cochin University of Science and Technology, Kochi - 682022, Kerala., India.
E-mail: sankaran.p.g@gmail.com

Deepa V. G., Department of Mathematics, Sree Krishna College, Guruvayur - 680104, Kerala., India.
E-mail: skcdeepa@gamil.com

Maya S. S., Department of Statistics, Maharaja?s College, Ernakulam - 682011, Kerala., India.
E-mail: ssmaya@gamil.com