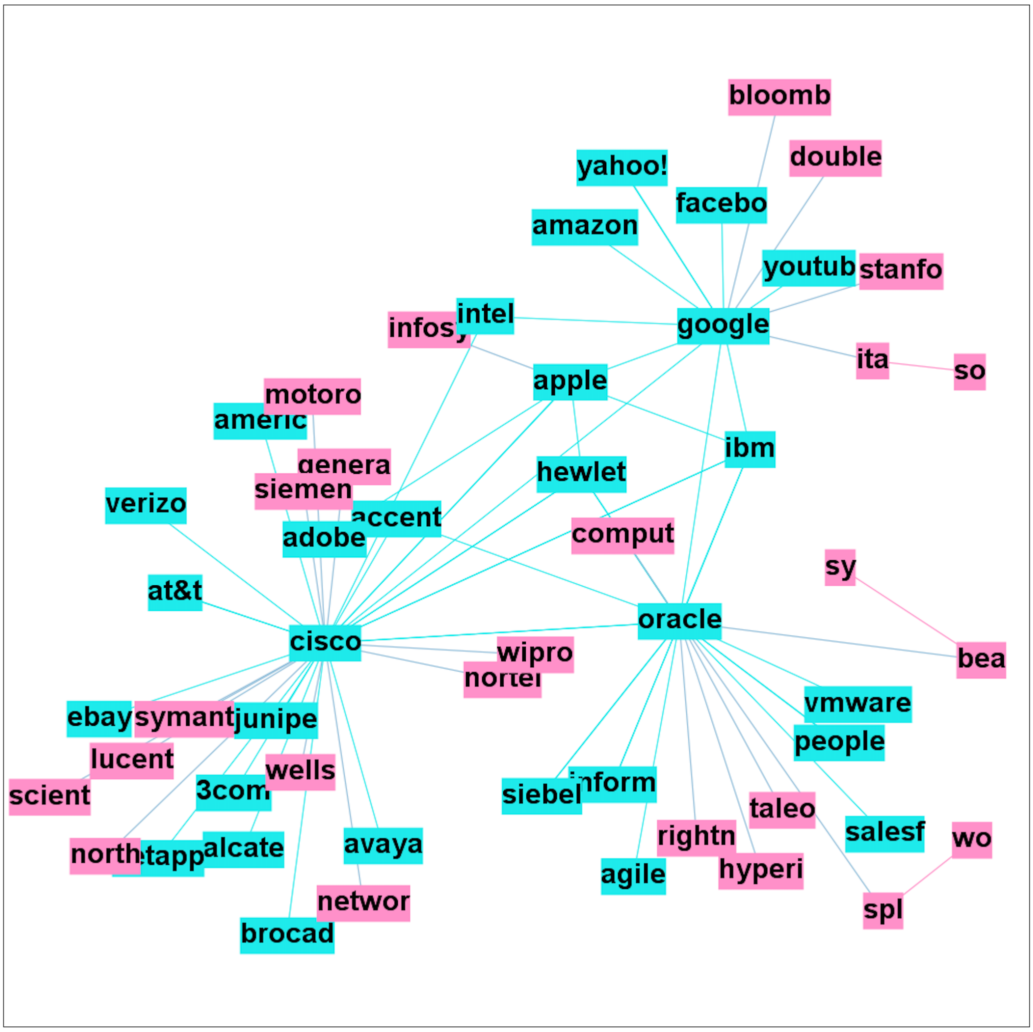
IS-5126 HANDS ON WITH BUSINESS ANALYTICS

HIRE FOR A LONG HAUL



Project Report - Group - 06

Jayakumar Alagappan Meenakshi	A0134431U
Krishna Kannan	A0134475A
Raghavendhra Balaraman	A0123443R
Vishnu Gowthem Thangaraj	A0134525L

1. INTRODUCTION

The hiring process in an organization is key to identifying the right employees by the employer. While the existing processes varies for different employers, often common measures are set in place to ensure the right employee is hired based on the requirements and skill set. Several rounds of technical interviews, HR interviews, stress interviews and what not. However, certain key factors still remain hidden and are not explicitly transparent to the employer.

“If hired, how long will the employee stay in the organization?”

“How and why have past employees of the organization moved away from the company?”

“Are my rivals intentionally or inadvertently stealing my best employees?”

We aim to provide insights and key findings that would help employer assess best candidate who fits the position based on his longevity in the company.

Why Should You Care ?

There are often scenarios where the employers can do very little about it :

- Employer conducts rounds of technical interviews and HR interview, practices which are perfected over the years to find the best employee based on the context of skill set and requirements. (Recruitment Costs)
- Employer provides training for the hired employee and develops the employee for the role he was hired for. (Training & Development Costs)
- Employer finally deploys the employee in a project and is able to see investment (Employee Compensation) be turned into productivity. (Administration Costs)

- The Employee then decides to leave the organization.

Employee Attrition is a very critical and inevitable problem that any organization faces, but absolutely want to keep the rate to a low minimum as possible, especially in the IT industry. A high attrition rate also causes a host of other problems to the employer, apart from the above incurred costs. A high attrition rate also inversely corresponds to the trust value and integrity of the company.

Herd Behaviour in Employee movement patterns

Also we could see examples of the 'herd behaviour' when it comes to employee attrition and movement patterns. An employee leaving the organization (for some reason X) has an effect on his colleagues and batch mates, who tend to consequently leave the organization as well (maybe not for the same reason X) because of herd behaviour.

Proposed Solution

We would like to provide solutions that would analytically help the employer identify potentially 'best fit' candidates, not just for their skill set and the requirements of the position but also for the expected longevity from the employee. We considered hiring factors which could lower the attrition rate of the organization. We would like to form the employee fitness score by considering various measures of the employee such as Experience, Degrees obtained, Schools attended, the employee's current employer's ratings such as culture, compensation, work-life balance and other ratings and various measures of the employer such as the history and characteristics of similar past employees in the same position.

We also considered employee movement patterns to get an understanding of how employees switch companies and how this could potentially help improve the organization's retention measures.

Identification of Key Factors

We also conceptualized factors involving both employees-employers, a unique measure that truly describes elements in the real world. For example, there could be a potential employee and potential employer who could have a high best-fit score if they come together, when only the employee-employer parameters are considered. We carefully crafted "Improvement parameters" to bring these elements into our model. These improvement measures are both indicators of change and severity for the employee employer combination.We also considered location metrics such as the "timetocommute" for the employee while forming the model. Current market models for prediction of stay are largely either employee factors or employer factors, we believe the inclusion of "elements" which are combination of both employee-employer characteristics in our models give us the 'edge' for a more accurate prediction.

2. TOOLS USED

Purpose	Tool
Web Scrapping	Python 2.7, Java
Statistical Analysis	R 3.1.3
Database	SQLite 3.8.2
Data Visualization Tools	R, Cytoscape 3.2.1, Gephi 0.8.2, NodeXL, Tableau
Framework	Circulo

3. DATA PREPARATION

3.1 Data Source

The data collection for this project largely comes from 3 main sources which includes LinkedIn, Glassdoor, Crunchbase. The below table highlights some important details about the data collection phase.

	Source	API	Language	Format
Employee Data	LinkedIn	LinkedIn REST API	Python 2.7	JSON
Employer Data	Glassdoor	Glassdoor REST API	Java	JSON
Employer Data	Crunchbase	CrunchBase V2 API	Python 2.7	JSON
Location Data	http://infoplease.com/us/census/data/	Static Data	-	Text

3.2 Extraction and Transformation of Data Sources

LinkedIn

The REST API resulting in JSON formatted files are converted to CSV files using Python scripts which resulted in three categories of linkedin data i.e., Employee's Personal data, Education data and Employment data. The CSV files are then loaded into tables as shown in the figure.

Glassdoor

Glassdoor community offers a variety of ratings about the companies which helps to learn about the employers. This has been leveraged on to arrive at a score for every company. Glassdoor provides a lightweight API that responds to the http request with JSON which contains detailed information about the company. First the list of companies for which ratings are required is collected, parsed and sent as a request and obtained rating information from the API.

CrunchBase

The CrunchBase data set obtained from the API are loaded in to database tables with the fields shown in the figure.

Location

The location details which includes ‘time to commute’ and ‘number of similar industries’ are obtained from Infoplease.com . The static data are then scraped into appropriate CSV files which are then loaded into SQLITE tables.

3.3 Data Preprocessing

3.3.1 Removing Outliers

In the first phase of data preprocessing, outliers were identified and removed. The following are some of the outliers that were removed during this phase

- HTML contents in the data sets
- Missing/NULL values in some of the important fields of the data set which includes, employee-title, company name, company ratings etc.,
- Records containing Non-ASCII characters

3.3.2 Extracting Data Subsets

As the scope of this project is restricted to US and Singapore based companies, the data set is filtered on all United States and Singapore based employees and employers. In addition, the dataset is further subjected to industry based filtering. Since the analysis involves identifying the migration pattern and the expected duration of the employees in Information Technology domain, the dataset is aggregated on the employees and employers who belong to IT & Services, Computer Hardware and Software, Telecom, Security, Internet, Networking and Data Management based industries. For location based filtering, we have identified 75 areas in US where employee concentration in Software and IT related services is dominant.

3.4 Data Mashup

After cleaning and preprocessing of dataset obtained from various sources, the data is now mashed to form a single source using which the models are built for further analysis. During this mashup phase some of the attributes are being added to help analyse the data better.

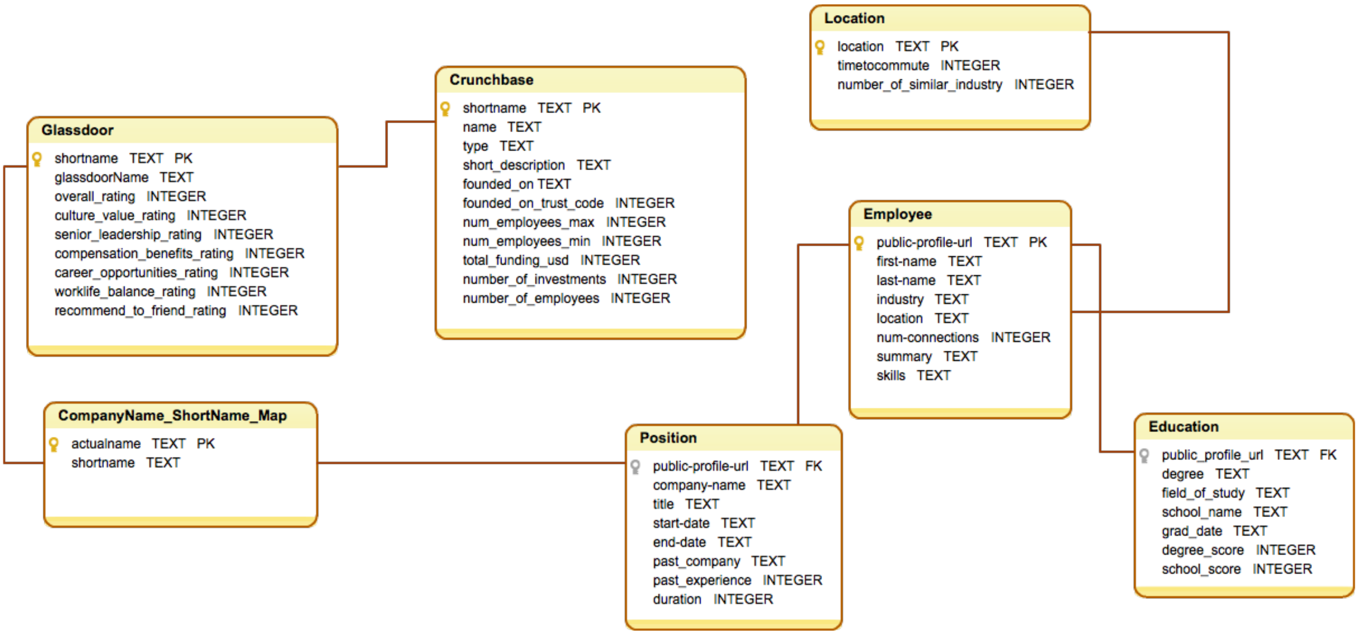
School Score - A static dataset containing the list of all universities and their corresponding world rankings are obtained. These data sets are then matched against the educational information of the employees from the linkedin datasets to identify their school rankings. The ranking are then normalized to a scale of 5 and are named as 'schoolscore'.

Degree Score - This attribute is again computed in a way similar to school score. From the linkedin education dataset, a list of unique degrees are identified. A score against each degree is added based on a keyword match. For ex., employees who hold a 'Master' degree which includes MS, Msc, MTech, MComp, ME etc are given a higher score. These score are later normalized to a scale of 5 and are named as 'degreescore'.

LinkedIn Actual Name - Short Name Map : Since the LinkedIn's 'company_name' field is a textbox, we found lot of redundant names given for a single company. For ex., a company like 'HP' can be written as 'HP', 'H P', 'Hewlette-Packard', 'Hewlette Packard', 'Hewlette

Packard(HP)’. Although all the variants are correct, but it makes searching difficult and creates redundant data. In order to remove such redundant entries, some common names (like a short name of the company) are assigned. Thus, from our example above, all the variants of Hewlette-Packard are mapped to a common name ‘HP’. This map is built by creating a table named ‘LinkedIn Actual Name - Short Name Map’ which maps the actual company names with their corresponding short names. These short names are then used to connect to glassdoor and crunchbase datasets.

The 4 distinct data sources i.e., LinkedIn, Glassdoor, Crunchbase, Location are now assigned with appropriate primary and foreign keys to help join the datasets without any data loss. The final schema design of the 4 data sources mashed up to form one data set is as shown below.



4. MODEL BUILDING

4.1 Regression Modelling

Attribute Selection : The analysis primarily focuses on predicting the expected duration of a candidate in a company for a particular position. The basic idea is that apart from employee based factors and employer based factors, the duration of stay by any candidate largely depends on employee-employer factor i.e., how well he fits in that particular company. The following parameters has been considered important and plays a major role in an employee's stay.

Employee Factors : pastexperience, degreescore, schoolscore, culture_employee, compensation_employee, career_employee, worklife_employee

We have assumed that an employee's current company's culture, compensation, and work-life balance ratings reflects significantly on the employee himself. As a result, his/her current company ratings are used as an employees key factor.   These factors are named as 'culture_employee', 'compensation_employee', 'career_employee', 'worklife_employee'.

Employer Factors: overall_rating, culture_rating, leadership_rating, compensation_rating, career_rating, recommend_to_friend _rating

Employee-Employer Factor: The relationship between the employee and employer factor is identified from the below formula.

(rating of the new company - rating of the current company) * (6 + rating of the new company)

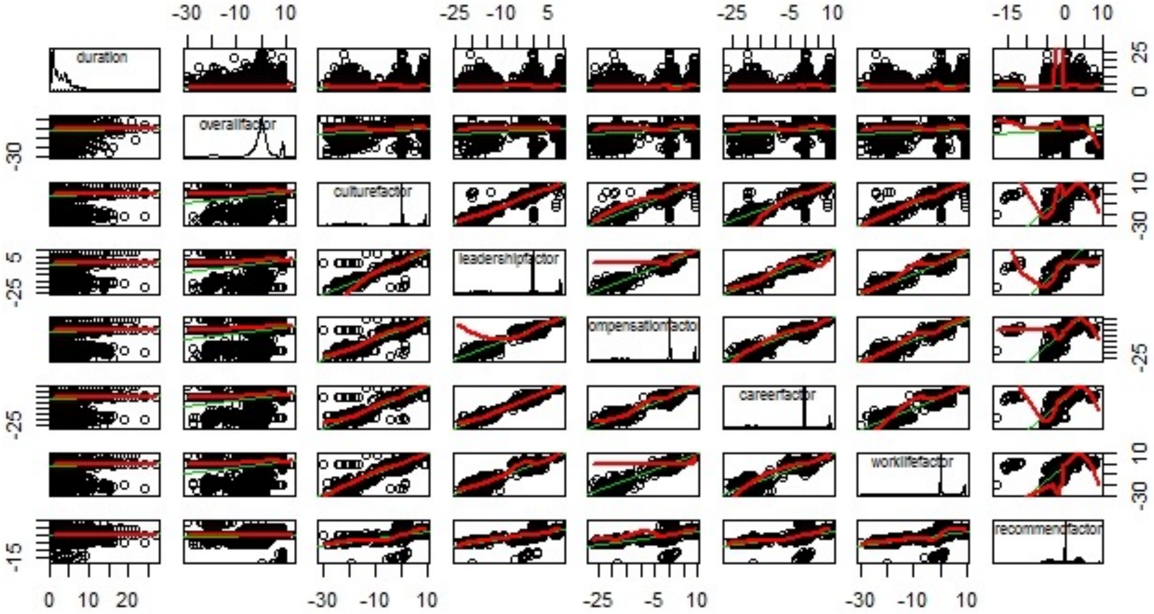
The below table shows few examples of how the above formula is calculated for different ratings

Rating	New Company	Current Company	Formula	Score
Overall	2	1	(2-1)*(6+2)	8
Worklife	1	2	(1-2)*(6+1)	-7
Compensation	5	4	(5-4)*(6+5)	11

Location : time_to_commute, number_similar_industries

Linear Equation Model :

After identifying the factors that affect the employee stay in a company, we have aggregated those parameters that are more related to the duration of employee in company using scatter plot functions in R. The below is the scatterplot diagram of some of the significant parameters.



The above scatter plot matrix is obtained for some of the Employee-Employer parameters such as Culturefactor, Leadershipfactor, Worklifefactor, compensationfactor, careeropportunity factor, recommendtofriendfactor. From the above scatter plot diagram, we can see that the parameters except recommendtofriendfactor looks correlated to the duration of stay by an employee. From these preliminary analysis, linear regression model is built with ‘duration’ as the target variable and rest of the above parameters as the predictors. The below equation summarizes the linear model,

Expected_Duration : **β**0 + **β**1pastexperience + **β**2degreescore + **β**3schoolscore + **β**4compensation_employee + **β**5worklife_employee + **β**6overallrating + **β**7culturerating + **β**8worklife + **β**9culturefactor + **β**10leadershipfactor + **β**11careerfactor + **β**11worklifefactor + **β**12duration_of_employee_in_same_position + **β**13time_to_commute

4.2 LinkedIn Graph Construction

The vertices of the graph represent the companies in the network. The directed edges represent the kineticism i.e. movement of employees between the companies. The information shown beneath is used to build the edge weights.

1) Mean term i.e. duration in a designation held by employees in a company.

2) Total number of employees switching from one organization to another.

Edge construction

Verbalize there subsists an edge between the companies A and B i.e. there has been employee switches between A and B,

Directed Graph

The number of switches from A to B/mean term in a designation held by employees switching from A to B is appointed as the edge weight.

Community Detection

Community detection is a noteworthy part of network examination. The goal is to identify how vertices i.e. companies in the graph should be assembled into communities/clusters. It is the issue of relegating the vertices of the graph into subsets such that the vertices belonging to a subset are all related. We used Infomap, a community detection algorithm for directed graphs to identify the clusters in the network.

Trend Analysis

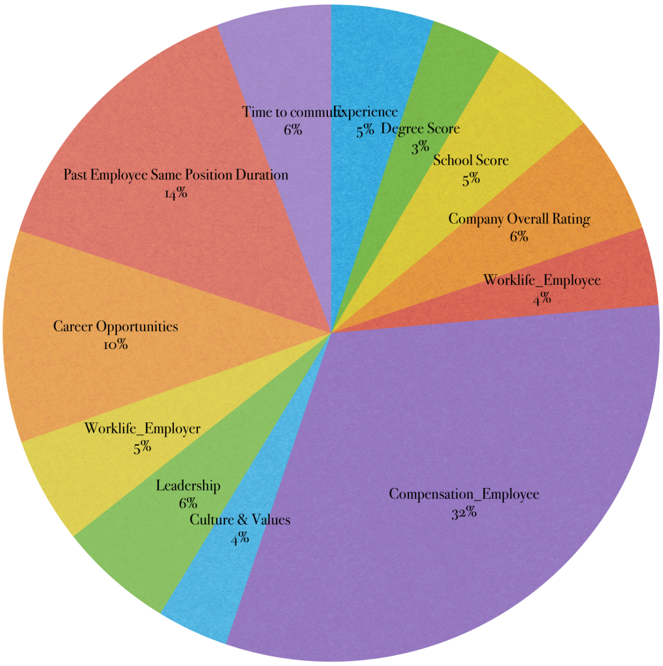
For a specific geology we recognize the movement of employees starting with one industry then onto the next, and movement starting with one company then onto the next inside a specific industry.

The inspiration to recognize this development is to identify the association between specific businesses and recognize how nearly certain commercial ventures are connected. Additionally the development inside an industry will help us to distinguish the most noticeable/preferred organizations in that industry.

5. DATA VISUALIZATION AND KEY INSIGHTS

From the linear regression equation obtained from the regression model, we consider the coefficient of the parameter in calculating their individual percentages. These percentages reflect how much each parameter plays an important role in predicting the stay of an employee in a company.

After analysing the migration pattern and the regression model performed in the previous step, the following are some of the insights we would like to provide,



- Individuals who stay in an organization longer than 5 year get paid 40% or less
- For the most part, we can see that individuals with 2 years of experience have the most astounding leaving rate from any organization
- Symantec, Walls Fargo, Wipro, Apple, 3COM, Microsoft, Intel, Oracle and HP are the targeted firms from where most of the people move to cisco.

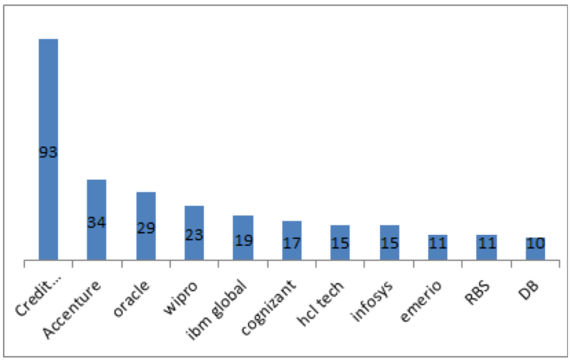
Trend Analysis

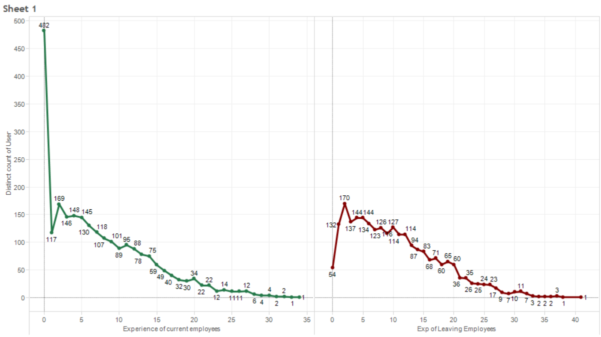
We ran the Infomap against US/SG - IT data sets. We took the top communities from each data set and analysed the trend pattern.

Singapore Data Set

Top Community 1 utilizing number of switches as weight:

The top community of Singapore is composed of Financial services companies with Citi bank and Credit Suisse having the most astounding In degree and Accenture with the most





The trends of distinct count of User for Experience and Exp Leaving. The data is filtered on Exp Leaving, which keeps 40 of 58 members.

astounding Out degree. The greatest betweenness centrality is for Accenture, trailed by Cognizant, Citibank and Barclays.

Top Community 1 utilizing number of years of experience as weight:

When we investigated the group utilizing number of years of experience as weight, for individuals with under 10 years of experience we discovered that Credit Suisse is the most coveted organization took after by Accenture and Oracle.

US Data Set:

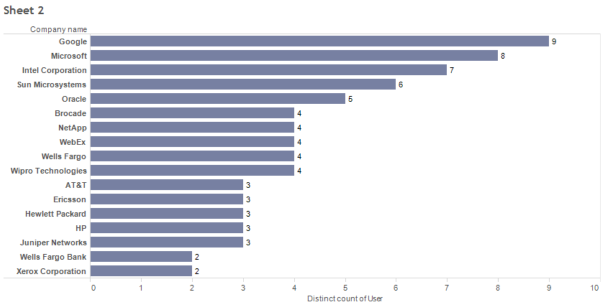
Top Community 1 utilizing number of switches as weight:

This group comprises of web, administration and item based innovation organizations commanded by Microsoft regarding degree centrality. After Microsoft the organization with most astounding centrality is Cisco trailed by Google, Oracle and HP each with a centrality of 12.

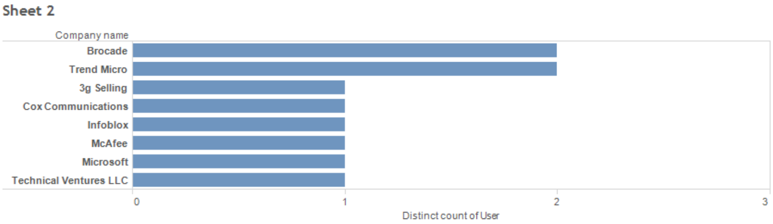
The main 5 organizations regarding betweenness centrality are as per the following:

Cisco: 717 Microsoft: 613 HP : 439 Oracle: 403 Infosys: 356

The most extreme movement of representatives were from Microsoft to Amazon and Microsoft to Google (81 and 75 respectively) which are the third and fourth most noteworthy developments in the group separately.



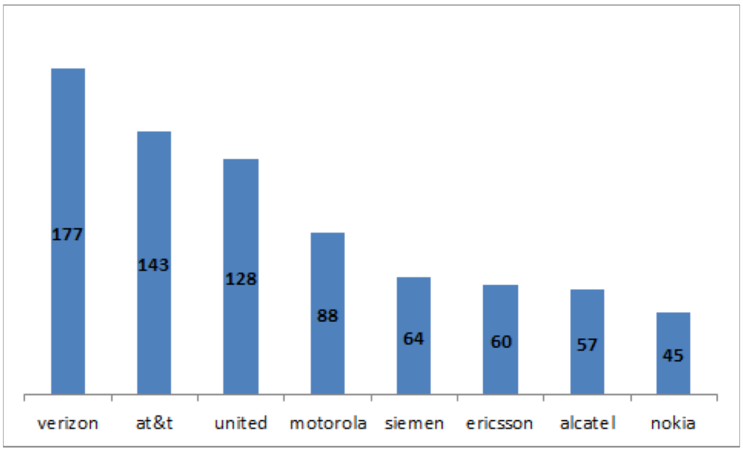
Distinct count of User for each Company name. The data is filtered on Experience, which keeps 27 of 58 members. The view is filtered on Company name, which keeps 1,087 of 1,489 members.



Distinct count of User for each Company name. The data is filtered on Experience, which keeps 30, 31, 32, 33 and 34. The view is filtered on Company name, which keeps 1,088 of 1,489 members.

Top Community 1 utilizing number of years of experience as weight:

When we dissected the group utilizing the duration as weight, we discovered that for people having background under 10 years, Verizon is the most craved organization. Likewise the main 10 most coveted organizations with both the weight apportions swings to be same.



6. CONCLUSION

Our proposition will prove to be a novel method of determining whether a hiring employee will be staying with the employer or leaving soon. As mentioned, we can also factor in specific parameters in our model for specific results, thereby ensuring its a win-win situation for all the concerned stakeholders. (For both the employee and the employer) Movement patterns analysis helps us identify what happens intendedly and inadvertently as well. It would help the employer evaluate their own methods, evaluate their own actions and conclude what changes are needed to improve their retention rate.

There are varied applications for the results. From a hiring employers perspective, a trend of how employees similar to the current candidate being hired have faired in their organization. Their tenure, their movement patterns from the organization would be identified. Although this could have adverse effects of stereotyping a potential candidate wrongly, it is relatively negligible and largely successful.

From a employee's perspective, a trend of how he/she should switch companies in order to reach their dream company can be identified. This could literally be "the shortcut" that people need to achieving their dream company.

We could extend our project to involve psychometric analysis datasets (datasets with survey results maybe, or other datasets that captured thoughts) in future to see how people of say 2 years experienced think which factor is important to switching (some might think work-life balance, culture and so on). This could be a key factor, as even before the entire interview process, recruiters would know how an employee of that situation is more likely to think.