Comprehensive Exam Answers

By Dieudonne Ouedraogo

2/1/2018

PART1 ——DR.HIROKI SAYAMA———-

S1

Consumer Behavior Analysis In Modern Environment

Abstract

Social networks activities are parts of our lives, and we interact with others through those means more often. As the number of users grows and technology improve, a new reality is taking place: even though most people identify themselves as unique, they tend to show preferences that could be clustered into groups. We are more willing to share our information on the social network; one may argue that we are becoming more predictable than before. We can defined groups or communities by labeling preferences and characteristics as instances where an individual is using perspectives or values as the basis for his or her current behavior. A group could guide the action in some situations.Those groups could be segmented as well.

Within groups either directly or indirectly, studies showed that more experienced members serve as experts and leaders and newer members seek advice and information directly or indirectly as well.We could argue that the collective behavior of the entire network influences any member. Regarding investments being made by companies on social media marketing, it is now indisputable that great insights could be extracted from social networks to drive business decisions and gain competitive advantage. Recent studies in network science reveal the presence of well-defined structures in social networks; an example is the presence of homophily which shows individuals with similarity tending to connect to each other. This paper examines the structural qualities of Social Networks towards the identification of trends in consumers behavior, it gives insights to the characterization of consumer behavior, particularly in the area of predictive analytics.

Keywords:

Consumer Behavior Analysis, Network Science, Data Mining, Trend Discover, Predictive Analytics, Personalization

Literature Review

Twitter has been extensively used on predicting trends in many research because of the relatively small size of its attributes. In a broader sense of social network being used to forecast events, Achrekar et al.[1] used Twitter to predict the trend of the flu virus. They successfully used auto-regression models on tweets to accurately predict the numbers published by the Center for Disease Control (CDC).While the CDC wait to collect actual cases to generate figures, their model could quickly predict outbreak and could be used to save lives.Iyengar et al. were able to predict the start and the end of a set of sports, weather and social activities using SVM classifier and hidden Markov Model on Twitter data.[2] Peng et al. investigated re-tweets patterns using conditional random fields.They defined features as the content influence, network influence and temporal decay factor.The results showed that re-tweet predictions could be substantially improved under social relationships compare to a baseline environment[3]. Gloor, Nann,and Schoder used structural qualities to find betweenness centrality of actors by weighing the context of their positions in the network, they successfully

predict long-term trends on the popularity of movies and politicians[4]. Understanding the underline structure of social networks and their relationship to each other is vital on predicting the behavior of nodes, in that sense, Mislove et al. studied the structure of different online social networks. Their results confirm the presence of power-law, small-world, and scale-free properties of online social networks they observe that the in-degree of user nodes tends to match the out-degree[5]. Cantonese et al. analyzed the properties of social networking graphs.They examined scaling laws distribution of friendship and centrality measurements [6]. A useful tool in consumer behavior prediction is "Link Mining". Algorithms using this technique are designed to support performance among some activities including question answering, information retrieval and web-based data warehousing [7]. Erbs et al. proved that training data and data volume improve performance in link discovery with text-based approaches[8]. Qian et al. used link mining techniques on the Enron mail corpus data and were able to show communities within linked nodes, they were able also to identify 'common friends' using cluster analysis.[9]. Other methods explored link-predictions with applications for exploring data, distributed environments, and spam analysis. [10][11][12]. Research in Social Networks is also visible on current search technology including Page Rank and HITS [13][14][15]. Using these techniques, Bharat, Henzinger, and Chakraborti presented variations that utilize web page context to weight pages and links based on relevance.[16][17].Sugiyama et al. used the topological structure of a graph to successfully combine few methods including network, quantitative, semantic, data processing, conversion and visualization-based components [18]. Research in Semantic Web technologies also yielded development in Social Networks. In that sense Zhou, Chen, and Yu combined an ontology-based Social Network along with a statistical learning method towards Semantic Web data using an extended FOAF (friend-of-a-friend) ontology applied as a mediation schema to integrate Social Networks and a hybrid entity reconciliation method to resolve entities of different data sources [19]. Thushar and Thilagam used Semantic Web technology for the identification of associations between multiple domains within a Social Network [20]. Several Relational Learning methods have supported Social Network analysis predicated on the concept of homophily-based associations to support learning. In that context, we have the application of probabilistic modeling [21] collaborative relationship [22] and inference-based approaches [23]. Visualization techniques are being used, and it substantially helps in studying Social Networks dynamics. Batajelj and Mrvar created tools for the visualization of large-scale networks where it is possible to identify vertices and relations between clusters [24]. Noel et al. calculated inter-item distances among combinations of elements from which hierarchical clustering dendrograms are visualized to enhance measurement consistency between clusters and frequent item-sets. They introduced an application of association mining to the visualization of link structures. Important frequently occurring higher-order item-sets are often obscured by the poor pairwise treatment of traditional analysis. The approach they take here involves the discovery of frequently occurring item-sets of arbitrary cardinalities, and the assigning of importance to them according to their support frequencies[25]. Levng et al. created Social Viz which provided users with a means to view frequency relationships among multiple entities in a network [26]. They used frequent pattern mining and visualization techniques. a visualizer called SocialViz is developed for providing users with frequency information on the social relationship among multiple entities in the networks. SocialViz could be used a standalone visualization tool, or as an additional tool to existing visualizers, for social networks exploration[26]. David Alfred et al. used a collection of twitter message to extract metrics that determine the effect of key players and find a correlation between their graph structure and the market share of three primary mobile Operating System[27]. Sharad Goel, and Daniel Goldstein used retail data and applied logistic regression and five-fold cross-validation to compute the likelihood of an individual making a purchase based on his contacts past activities.The results show that individuals with contacts who made a purchase before are more likely to purchase than individuals with connections who did not have any previous purchase[28]. Yoon et al. used S&P companies data from 2010-2015 and math them to 24 million user comments directed at those companies' Facebook posts.They tested hypotheses using fixed effect(FE) and random effects(RE) and dynamic (generalized method of moments) and reached to the conclusion that digital engagement volume has significant positive impact on revenue[29] John et al. cautioned the translation of "liking" on social media into an indicator of an intent to make a purchase.Through their study they discover based on more than 14000 cases, that more features in addition to the button "Like" are needed to make the accurate prediction on the purchase[30]. Chong et al. conducted an experimental study on the consumer engagement behavior(CEB) and were able to show using ordinary least square (OLS) regression models that Facebook and YouTube activities positively correlate with box-office revenue, however, their results are not conclusive for twitter.They proposed and tested a set of metrics[31]. Ding et al. used

Pre-released movies "likes" data from Facebook and discovered that more campaign on those Pre-released data increases revenue for a film [32]. Yung et al. propose an experimental model where businesses can target new customers; when a customer visits a store, recommendations are guided using the client social media data. The preliminary results show that companies can generate substantial revenue utilizing this process[33]. Hyunmi et al. investigated the ewom(electronic word-of-mouth) of different social networks using Roger's innovation diffusion model on collected daily data from movies and the results pertain that Twitter influence on box office revenue is more significant in the initial opening stage[34].

Limits of the previous approaches.

Most of the work elaborated above even though spread around different techniques do handle Social Network analysis in a static fashion where nodes or actors dynamics are not taken into account. The arrival or departure time of agents are not taking into consideration, but we believe those features can lead to better insights.The geospatial distribution of the contacts is somehow neglected in the studies.Very often our contacts on networks are spread around the globe and depending on the geographical position, reality and culture can create a barrier that is hard to overcome so we believe a segmented approach can be helpful. Most of the work above also are somehow single network an approach, individual may have preferred using different social networks, and we believe an approach that combines multiple networks analysis may be more conclusive. A handicap with most of those research is the limited scope of Social Networks analysis only.Most of the data available in Social Network are unstructured, a hybrid approach where also structured transaction data are used can lead to better predictions.If a new node joins a network, we can make a recommendation based on his contacts preference, but an old node which has transaction data available in retailer database could have a better recommendation based on both analysis.

Methodology

We defined a framework where nodes are individuals or users. Unlike previous studies, we consider the time factor when a node enter or exit the network. We define and track metrics that define nodes activities: number of comments or "like" and browsing time on social networks related to time.We define similarity measures as characteristics shared by different nodes, and we cluster nodes based on those similarities. We combine the use of multiple networks datasets. We describe a second set of nodes as products, and we specify characteristics that differentiate products. We study the proprieties of nodes or group of nodes as well as their edges(relationships).Unlike previous studies, we consider two situations on multiple Social Networks, a condition where consumer's past transactional data is available and a situation in the absence of transactional data. We propose hybrid models for purchase predictions in either situation. We measure the performance of the models proposed on the actual data.

References

[1] Achrekar, H.; Gandhe, A.; Lazarus, R.; Ssu-Hsin Yu; Benyuan Liu; Predicting Flu Trends using Twitter data Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on Publication Year: 2011 , Page(s): 702 - 707

[2] Peng, Huan-Kai; Zhu, Jiang; Piao, Dongzhen; Yan, Rong; Zhang, Ying; Retweet Modeling Using Conditional Random Fields Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on Publication Year: 2011, pp 336-343

[3] Iyengar, Akshaya; Finin, Tim; Joshi, Anupam; Content-Based Prediction of Temporal Boundaries for Events in Twitter Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom) Publication Year: 2011, pp 186-191

[4] Wasserman, Faust, "Social network analysis: methods and applications" (structural analysis in the social sciences), Cambridge University Press, Cambridge.

[5] Measurement and analysis of online social networks by Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, Bobby Bhattacharjee In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (2007), pp. 29-42,

[6] Cantonese, Salvadore, De Meo, Pasquale, Ferrara, Emilio, Fiumara, Giacomo, Provetti, Alessandro, Crawling Facebook for Social Network Analysis, WIMS'11 May 25-27, 2011 Sogndal Norway

[7] Knowledge Discovery and Retrieval on World Wide Web Using Web Structure Mining

Boddu, Sekhar Babu; Anne, V.P Krishna; Kurra, Rajesekhara Rao; Mishra, Durgesh Kumar; Mathematical/Analytical Modelling and Computer Simulation (AMS), 2010 Fourth Asia International Conference on, 2010, pp: 532-537

[8] Erbs, Nicolai, Zesch, Torsten, Gurevych, Iryna, Link Discovery: A Comprehensive Analysis, 2001 Fifth IEEE International Conference on Semantic Computing

[9] Acar, E.; Dunlavy, D.M.; Kolda, T.G.; Link Prediction on Evolving Data Using Matrix and Tensor Factorizations Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on 2009, pp 262-269

[10]Cai-Rong Yan; Jun-Yi Shen; Qin-Ke Peng; Ding Pan; Parallel Web mining for link prediction in cluster server Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on Volume: 4: 2005 , Page(s): 2291 - 2295 Vol. 4

[11] Caverlee, J.; Webb, S.; Ling Liu; Rouse, W.B.; A Parameterized Approach to Spam-Resilient Link Analysis of the Web Parallel and Distributed Systems, IEEE Transactions on Volume: 20, Issue: 10 2009, pp 1422-1438

[12] Rong Qian; Wei Zhang; Bingni Yang; Detect community structure from the Enron Email Corpus Based on Link Mining, Intelligent Systems Design, and Applications, 2006. ISDA '06. Sixth International Conference on Volume: 2Publication Year: 2006 , Page(s): 850 -855

[13] Web structure mining: an introduction da Costa, M.G., Jr.; Zhiguo Gong; Information Acquisition, 2005 IEEE International Conference on 2005

[14]L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the Web., Technical Report, Stanford Univesity, 1998

[15]NMF: Network Mining Framework Using Topological Structure of Complex Networks Sugiyama, K.; Ohsaki, H.; Imase, M.; Yagi, T.; Murayama, J.; Congress on Services Part II, 2008. SERVICES-2. IEEE Publication Year: 2008, pp 210-211

[16] J. Kleinburg, Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5): 604-632 1999

[17]K. Bharat , M.R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment. In ACM SIGIR International Conference on Research and Development in Information Retrieval, pages 104-111, 1998

[18]S. Chakrabarti, B.Dom, and P.Indyk Enhanced hypertext categorization using hyperlinks. In SIGMOD International Conference on Management of Data pp 307- 318, 1998

[19]Semantic Message Link Based Service Set Mining for Service Composition Anping Zhao; Xiaoyong Wang; Ke Ren; Yuhui Qiu;

Semantics, Knowledge and Grid, 2009. SKG 2009. Fifth International Conference on 2009 , Page(s): 338 - 341

[20]Thushar, A.K.; Thilagam, P.S.; An RDF Approach for Discovering the Relevant Semantic Associations in a Social Network Advanced Computing and Communications, 2008. ADCOM 2008. 16th International Conference on Publication Year: 2008, pp 214- 220

[21] Achim Rettinger Matthias Nickles, Volker Tresp Statistical Relational Learning with Formal Ontologies, ECML PKDD '09 Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II

[22] Kirsten, Mathias, Wrobel, Stefan, Inductive Logic Programming, Lecture Notes in Computer Science, 1998, Volume 1446/1998, 261-270, DOI: 10.1007/BFb0027330

[23] Chunying Zhou; Huajun Chen; Tong Yu; Learning a Probabilistic Semantic Model from Heterogeneous Social Networks for Relationship Identification Tools with Artificial Intelligence, 2008. ICTAI '08. 20th IEEE International Conference on Volume: 1

[24] Batagelj, Vladimir, Mrvar, Andrej, Pajek: Analysis and visualization of large networks, Graph Drawing Software Book. Junger, P. Mutzel, editors 2003

[25]Noel, S.; Raghavan, V.; Chu, C.-H H.H.; Visualizing association mining results through hierarchical clusters, Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on Publication Year: 2001, pp 425 - 432

[26]Leung, Carson Kai-Sang, Carmichael, Christopher L., Exploring Social Networks: A Frequent-Pattern Visualization Approach, IEEE International Conference on Social Computing, 2010

[27] David Alfred Ostrowski. "Social Network Analysis for Consumer Behavior Prediction". Accessible at: http://worldcomp-proceedings.com/proc/p2012/ICA3445.pdf

[28]Sharad Goel, Daniel C. Goldstein. "Predicting Individual Behavior with Social Networks"

[29] Attracting Comments: Digital Engagement Metrics on Facebook and Financial Performance Gunwoo Yoon, Cong Li, Yi (Grace) Ji, Michael North, Cheng Hong & Jiangmeng Liu Pages 1-14 | Received 28 Apr 2017, Accepted 07 Nov 2017, Published online: 24 Jan 2018 https://doi.org/10.1080/00913367.2017.1405753

[30]Leslie K. John, Oliver Emrich, Sunil Gupta, and Michael I. Norton (2017) Does "Liking" Lead to Loving? The Impact of Joining a Brand's Social Network on Marketing Outcomes. Journal of Marketing Research: February 2017, Vol. 54, No. 1, pp. 144-155. https://doi.org/10.1509/jmr.14.0237

[31]Chong Oh, Yaman Roumani, Joseph K. Nwankpa, Han-Fen Hu Beyond likes and tweets: Consumer engagement behavior and movie box office in social media Information & Management Volume 54, Issue 1, January 2017, Pages 25-37 10.1016/j.im.2016.03.004

[32]Chao Ding,Hsing Kenneth Cheng,Yang Duan,YongJin The power of the "like" button: The impact of social media on box office Decision Support Systems Volume 94, February 2017, Pages 77-84 https://doi.org/10.1016/j.dss.2016.11.002

[33]Yung-Ming Li, Lien-Fa Lin, Chun-Chih Ho A social route recommender mechanism for store shopping support Author links open overlay panel Decision Support Systems Volume 94, February 2017, Pages 97-108

https://doi.org/10.1016/j.dss.2016.11.004

[34]Hyunmi Baek,Sehwan Oh,Hee-DongYang,JoongHo Ahn Electronic word-of-mouth, box office revenue and social media Electronic Commerce Research and Applications Volume 22, March-April 2017, Pages 13-23 https://doi.org/10.1016/j.elerap.2017.02.001

S2.

The dataset could be downloaded from here

http://konect.uni-koblenz.de/networks/amazon-ratings

The dataset represents the ratings given by users on products during a specific period.The data was extracted from the website loaded into TextWrangler and converted into a text file.The data was then loaded into R studio.R programming is used to answer this part.The package igraph is used.To make conclusions and inferences more convincing, codes outputs and reports are embedded together!

Reading the data into the environment and loading necessary libraries

```
library(pander)## Printing libraries for display
library(knitr)## Printing libraries
library(igraph)## Network science library
##Load the data
ratings <- read.table("~/Downloads/amazon-ratings/amazon-ratings1.txt", quote="\"", comment.char="")
##Name the colomns
colnames(ratings)<-c("ID_from_Node","ID_to_Node","Edge_Weight","Timestamp")
###Let's display the first 10 rows
kable(head(ratings,10),caption="First 10 Rows of the data")
```

Table 1: First 10 Rows of the data

ID_from_Node	ID_to_Node	Edge_Weight	Timestamp
1	1	5	1117404000
1	2	1	1105916400
1	3	5	1105916400
1	4	1	1105570800
1	5	1	1104966000
1	6	5	1103497200
1	7	4	1081461600
1	8	5	1074985200
1	9	5	1071961200
1	10	1	1071788400

The data could be saved into a CSV file for easy future work; below is the code

```
write.csv(ratings,file = "Ratings.csv")
```

The Edge Weight which represents the values of ratings could be grouped to have a clear idea of the most rated product and the most active user (nodes)

```
d=ratings$Edge_Weight#Define a variable d which contain all the weights
td=table(d) #group d by distinctive values and assigned that table to variable
pander(td,caption="Ratings")
```

Table 2: Ratings

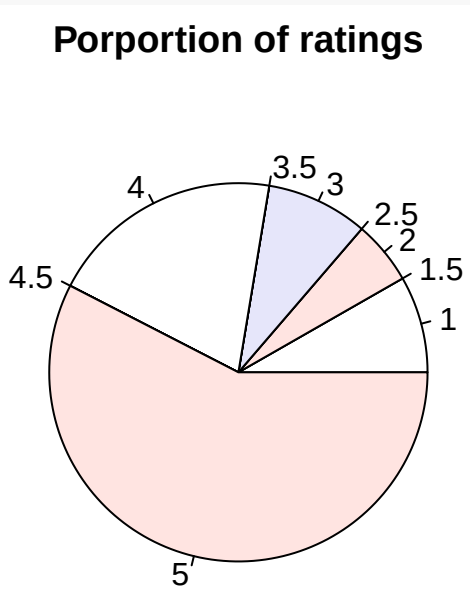
1	1.5	2	2.5	3	3.5	4	4.5	5
482809	17	316934	24	507386	76	1170161	211	3360423

Let's look at the most used rating value and the distribution of the ratings

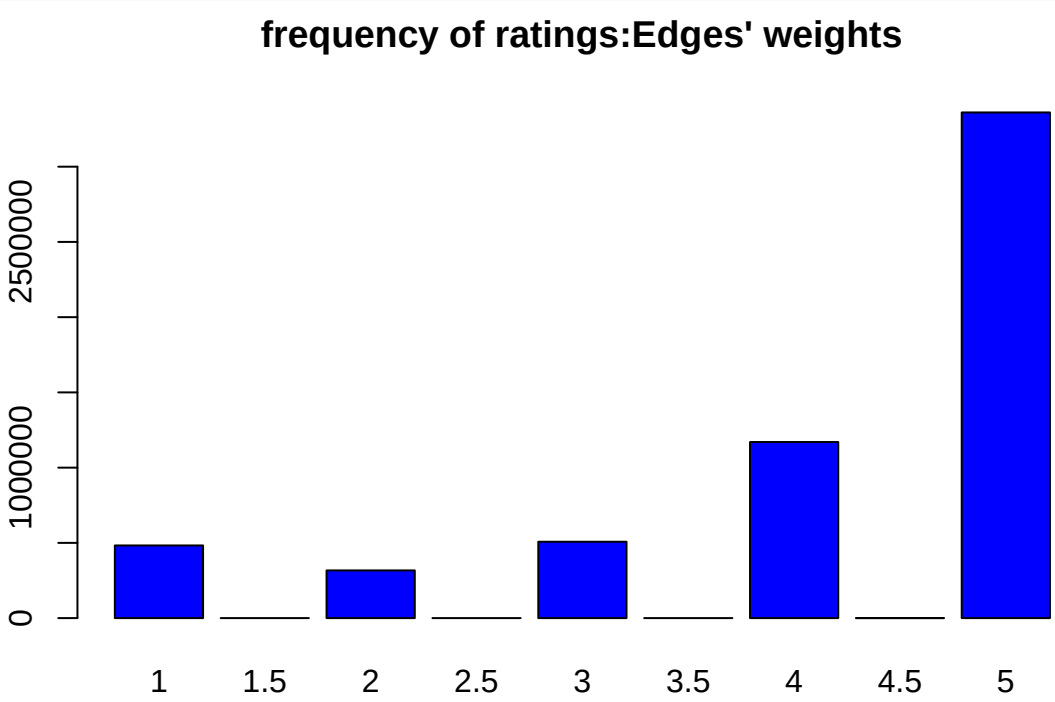
```
kable(paste("The most used rating appears",max(td),"times"))
```

The most used rating appears 3360423 times

```
pie(td,main="Porportion of ratings")
```




```
barplot(td,col="blue",main="frequency of ratings:Edges' weights")
```



Based on the above plot we can see the distribution of the ratings which indicates that it is not relevant to provide 1.5,2.5,3.5 rating choices: most people don't use them.

Let's define the most active user(node)

```
b=ratings$ID_from_Node# define a variable b and assign ID from nodes to it.
tt=table(b)# group by the same values
#Get the  most repeat value
kable(paste("The most active node has",max(tt),"ratings"))
```

The most active node has 12217 ratings

```
kable(paste("THE MOST ACTIVE NODE IS:",names(tt[tt==max(tt)])))
```

THE MOST ACTIVE NODE IS: 10662

The node '10662' has more links(ratings) to products than any other node in the network. This is the MOST ACTIVE NODE in network

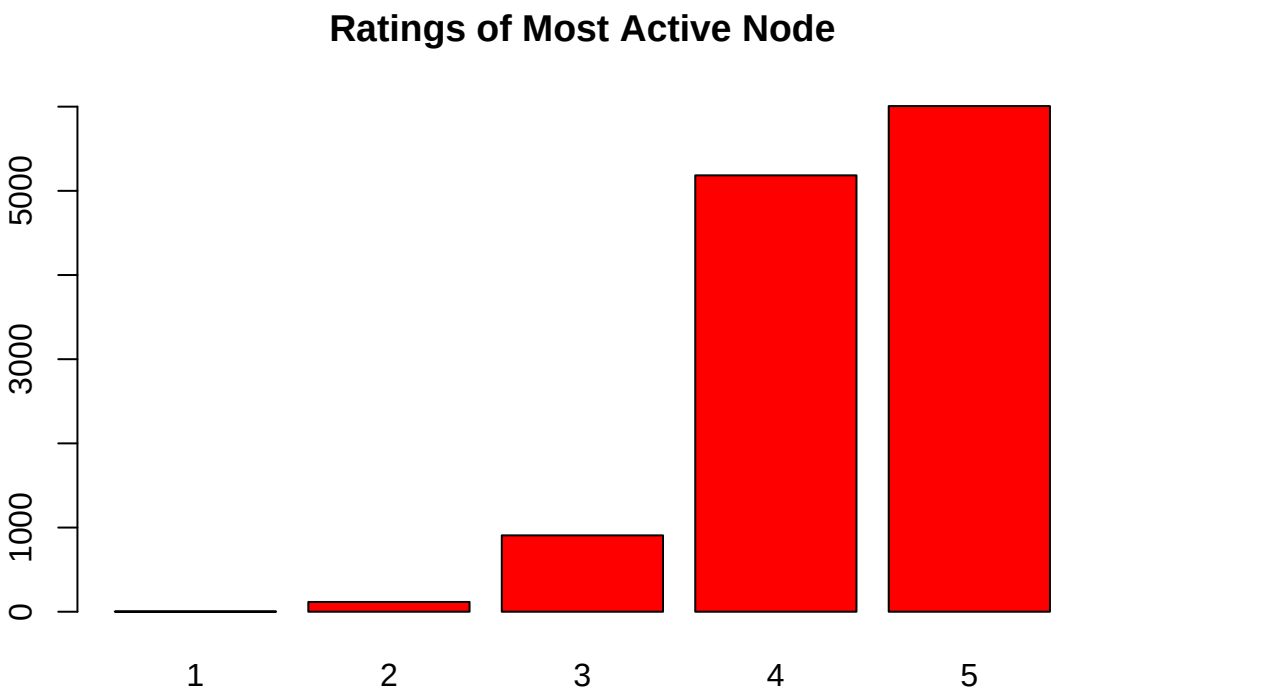
Let explore the distribution of his ratings.The data associated with this node could be saved into a file for future work(investigative work)

```
newdata <- ratings[ which(ratings$ID_from_Node=='10662'),]
#write.csv(newdata,"MostActiveNode.csv")
d=newdata$Edge_Weight#Define a variable d which contain all the weights
td=table(d) #group d by distinctive values and assigned that table to variable
pander(td,caption="Ratings of Most Active Node(User)")
```

Table 6: Ratings of Most Active Node(User)

1	2	3	4	5
3	116	907	5184	6007

```
barplot(td,col="red",main="Ratings of Most Active Node")
```



Node 10662 is the most active,but it may not be the node that provide the most positive rating!Positive ratings(5) can help drive more sales so more marketing directed to positive raters can be effective.

```
data5= ratings[which(ratings$Edge_Weight=='5'),]#Selecting only 5 ratings
data5_From=data5$ID_from_Node#assing a variable
t=table(data5_From)#Group by value
#max(t)
# the node below gave the most number of rating 5
#The node is
kable(paste("The Most Friendly User is ",names(t[t==max(t)])))
```

The Most Friendly User is 20898

"20898" gives the highest ratings than any other node,this could be refer as the most friendly node(user),his input help drive more sales as he makes the average ratings higher,in some environment he is called booster.

Now let's look at the product side

From the product standpoint, we can find the best-rated product.We are interested in products with '5' stars as a rating, and we want to find the product which gets the highest number of '5'.

```
#Finding ratings equal to 5 and finding to node value
data5= ratings[which(ratings$Edge_Weight=='5'), ]
data5_to=data5$ID_to_Node
t1=table(data5_to)
kable(paste("Maximum number of 5 received",max(t1)))
```

Maximum number of 5 received 2050

```
#2050 times a product has 5 ratings
kable(paste("Product with maximum ratings is ",names(t1[t1==max(t1)])))) #The name of the node
```

Product with maximum ratings is 1302

Product '1302' has more positive ratings than any other node. The characteristics of this product could be recorded and perceived as a benchmark for future products conception.This product could be label as "Amazon choice" as it is very well perceived and could be used to attract new customers.It could be label as the MOST LIKED PRODUCT in the network.

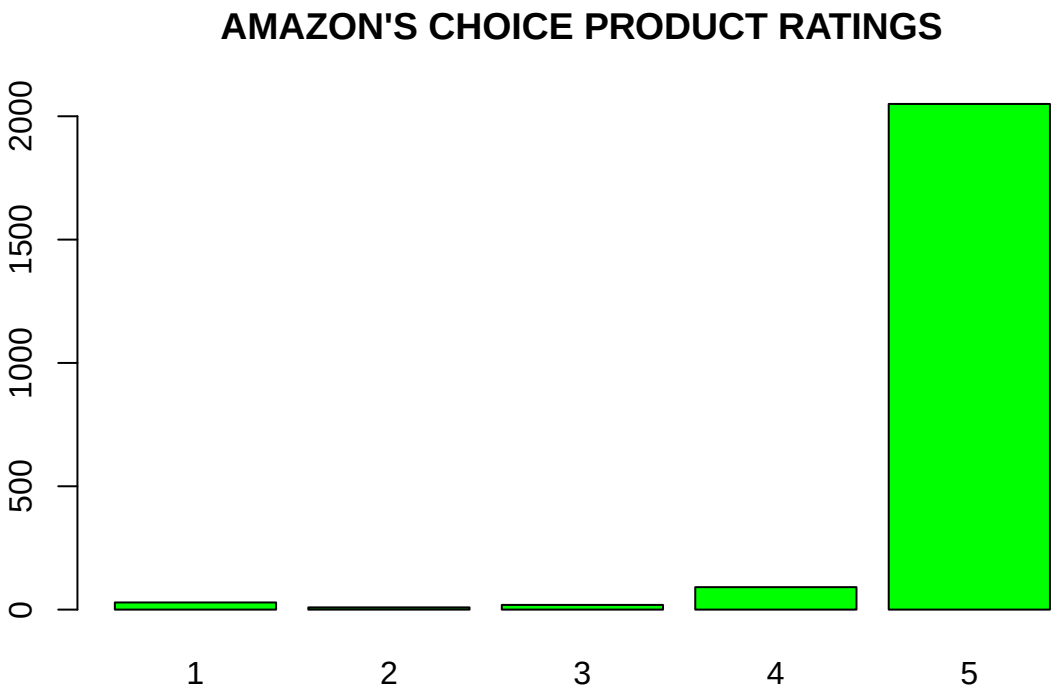
Let explore the distribution of its ratings.The data associated with this node could be saved into a file for future work (investigative work)

```
newdata2 <- ratings[ which(ratings$ID_to_Node=='1302'),]
#write.csv(newdata,"HIGHEST RATINGS.csv")
d=newdata2$Edge_Weight#Define a variable d which contain all the weights
td=table(d) #group d by distinctive values and assigned that table to variable
pander(td,caption="Most Liked Product ratings frequency")
```

Table 10: Most Liked Product ratings frequency

1	2	3	4	5
29	9	19	91	2050

```
barplot(td,col="green",main="AMAZON'S CHOICE PRODUCT RATINGS")
```



This plot indicates that few people are not happy with the product, so a closer targeted study with this group could reveal what needs to be improved in the product to make it more attractive to customers.

Let's transform the ratings data into a graph entity for further inferences

```
net=graph_from_data_frame(ratings,directed=FALSE)
class(net)#Let's check the type of net to make sure it is a graph
```

```
## [1] "igraph"
```

```
V(net)# Let's display few Nodes
```

```
## + 2146057/2146057 vertices, named, from f507b5e:
##     [1] 1     2     3     4     5     6     7     8     9     10    11
##    [12] 12    13    14    15    16    17    18    19    20    21    22
##    [23] 23    24    25    26    27    28    29    30    31    32    33
##    [34] 34    35    36    37    38    39    40    41    42    43    44
##    [45] 45    46    47    48    49    50    51    52    53    54    55
##    [56] 56    57    58    59    60    61    62    63    64    65    66
##    [67] 67    68    69    70    71    72    73    74    75    76    77
##    [78] 78    79    80    81    82    83    84    85    86    87    88
##    [89] 89    90    91    92    93    94    95    96    97    98    99
##   [100] 100   101   102   103   104   105   106   107   108   109   110
## + ... omitted several vertices
```

```
#E(net)#Edges
E(net)# Let's display few Edges
```

```
## + 5838041/5838041 edges from f507b5e (vertex names):
##    [1] 1--1   1--2   1--3   1--4   1--5   1--6   1--7   1--8   1--9   1--10
##   [11] 1--11  1--12  1--13  1--14  1--15  1--16  1--17  1--18  1--19  1--20
##   [21] 1--21  1--22  1--23  1--24  1--25  2--26  2--27  2--28  2--29  2--30
```

```
##  [31] 2--31  2--32  2--33  2--34  2--35  2--36  2--37  2--38  2--39  2--40
##  [41] 2--41  2--42  2--43  2--44  2--45  2--46  3--47  3--48  3--49  3--50
##  [51] 3--51  3--52  3--53  3--54  3--55  3--56  3--57  3--58  4--59  4--60
##  [61] 4--61  4--62  4--63  4--64  4--65  4--66  4--67  4--68  4--69  4--70
##  [71] 4--71  4--72  4--73  4--74  4--75  4--76  4--77  4--78  4--79  4--80
##  [81] 4--81  4--82  4--83  4--84  4--85  4--86  5--87  5--88  5--89  5--90
##  [91] 5--91  5--92  5--93  5--94  5--95  5--96  5--97  5--98  5--99  5--100
## + ... omitted several edges
```

Since this a giant graph where large computing power is required we will not be able to extract the characteristics directly on the graph,instead we sample 5000 observations and collect the metrics of the sample and make an inference on the initial graph

We set a seed at 100 for reproducible results

```
set.seed(100)
index<-sample(1:nrow(ratings),5000)# Here we index the sample items
Sampleratings=ratings[index, ]
#write.csv(Sampleratings, file = "NewRatings.csv")# This file could be used in Gephi
net=graph_from_data_frame(Sampleratings,directed=FALSE)
kable(paste("Mean Distance is :",round(mean_distance(net, directed=F),6)))
```

Mean Distance is : 1.099476

```
kable(paste("The graph density is:",round(graph.density(net,loop=FALSE),6)))
```

The graph density is: 0.000108

```
kable(paste("the shortest path is: ",max(shortest.paths(net,mode="all"))))
```

the shortest path is: Inf

```
#
kable(paste("The maximum Eccentricity",max(eccentricity(net,mode="all"))))
```

The maximum Eccentricity 3

```
deg <- degree(net, mode="all")
kable(paste("The diameter of this network is",diameter(net)))
```

The diameter of this network is 3

```
kable(paste("The maximum degree is:",max(deg)))
```

The maximum degree is: 12

```
kable(paste("The minimum degree is:",min(deg)))
```

The minimum degree is: 1

```
kable(paste("The Average degree is:",round(mean(deg),4)))
```

The Average degree is: 1.0376

```
Degree_Correlation=assortativity_degree(net,directed = F)
kable(paste("The degree correlation is:",round(Degree_Correlation,4)))# Degree correlation
```

The degree correlation is: -0.0255

The mean distance give us an indication of the average number of connections.

The Density is very low, indicating a low connection, the number of possible ratings is much higher than the actual ratings.This confirms that the distribution of ratings is skewed.

Shortest path, This value is infinity indicating that nodes are very distant in term of the number of edges to take to reach to them.

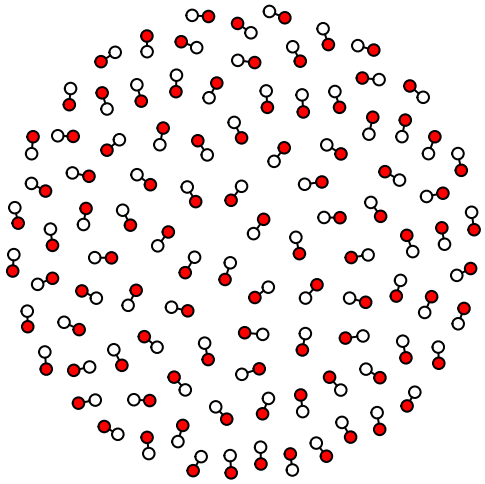
The degree correlation is negative since this value is negative it is an indication that the network is disassortative, meaning it is less likely that higher degree node attaches with higher degree node, it tends to attach to lower degree node!

The max,min and average degree;those numbers give us an indication of how many edges (the ratings in our case) are related to the node(user-product)

Sample Visualization

Here we will pick just 100 observations and vizualize a sample of the structure of the network

```
index<-sample(1:nrow(ratings),100)# Here we index the sample items
Sampleratings=ratings[index, ]
net=graph_from_data_frame(Sampleratings,directed=FALSE)
V(net)$color=ifelse(Sampleratings[V(net),2]==1,"blue","red")
plot.igraph(net,vertex.size=5,vertex.label=NA,edge.color="black",edge.width=E(net)$weight,vertex.color=
```



PART 2 —DR. HAROLD LEWIS—-

L-1 (Fuzzy Models)

Mandani-style fuzzy inference system

INSURANCE PREMIUMS

Actuaries at an insurance company decide to price premium from drivers based on the number of years he spent in school and their actual age.Prior internal data had shown somehow a relationship between those two inputs (age and education years) and the risk associated with the policyholder, so premium to be paid each month must reflect the risk associated! They believe premium between 200 and 1200 depending on the driver are reasonable to stay competitive and to avoid bankruptcy. so the output is

$Y_p = \{200, 400, 600, 800, 1000, 1200\}$;

Y_p = Premium

This set could broken down into L ,M,H respectively for low ,medium, and High premium.

The membership functions defining those values are TFN

$L : (-\infty, 200, 400)$

$M : (200, 600, 100)$

$H : (600, 1000, \infty)$

Ages

Drivers ages are between 16 and 60+ and defined by AG The set could be split into

Y,M,A respectively for Young,Middle age and Adult.

The membership functions are TFN

$Y : (-\infty, 20, 28)$

$M : (20, 28, 36)$

$A : (28, 36, \infty)$

Educations

Education is between 0 and 20+ The set could be split into L,M,H for Low,Medium and High The membership functions are TFN

$L : (-\infty, 8, 12)$

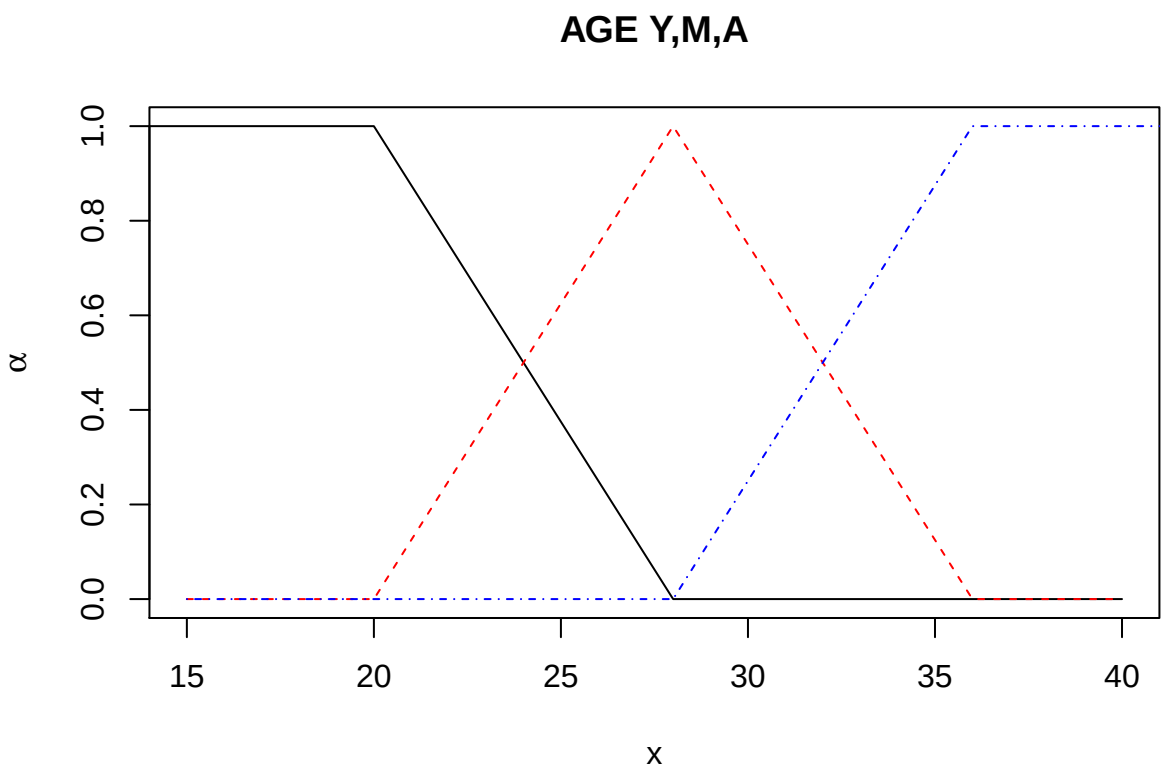
$M : (8, 12, 16)$

$H : (12, 16, \infty)$

Below I used the library FuzzyNumber from R to plot the membership functions.TFN is a particular case of TFRN with just a repeated value!So to get a TFN, we use Trapezoidal-FuzzyNumber function with a repeated sequence.

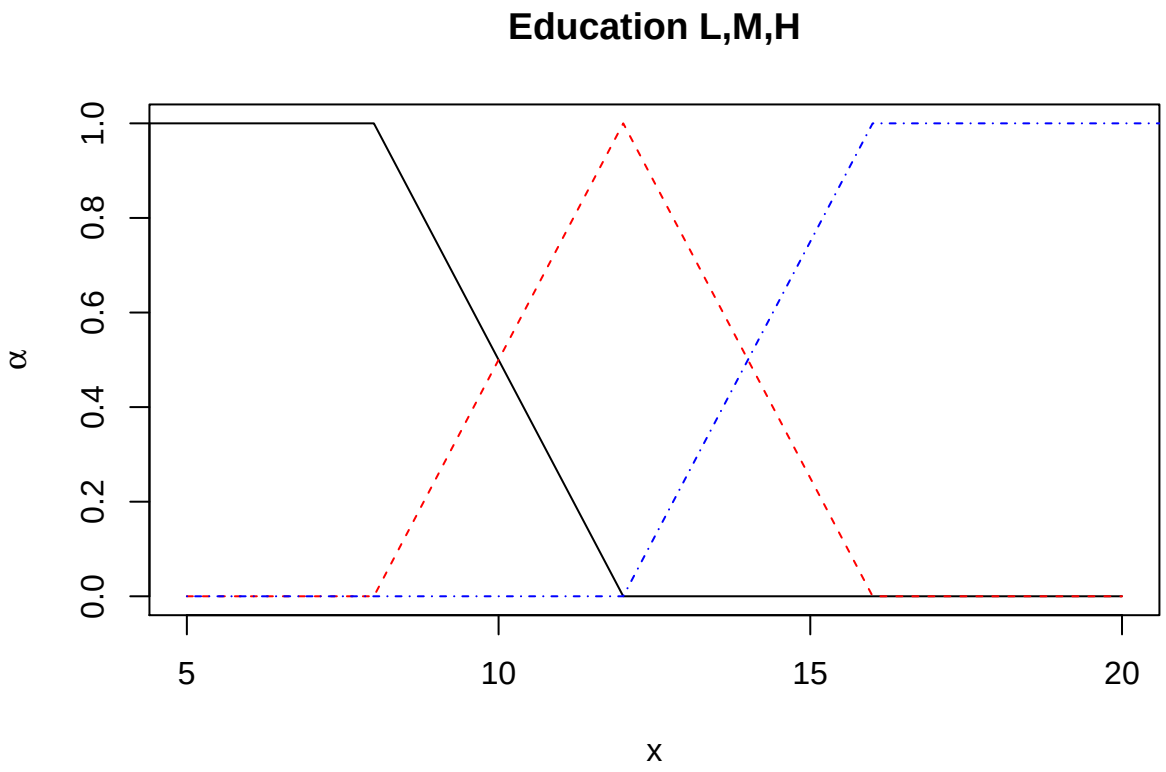
AGE

```
library(FuzzyNumbers)
Y <- TrapezoidalFuzzyNumber(-10^20,20,20,28)
M <- TrapezoidalFuzzyNumber(20,28,28,36)
A<-TrapezoidalFuzzyNumber(28,36,36,10^20)
plot(Y,xlim=c(15,40),main="AGE Y,M,A")
plot(M, add=TRUE, col=2, lty=2)
plot(A, add=TRUE, col=4, lty=4)
```



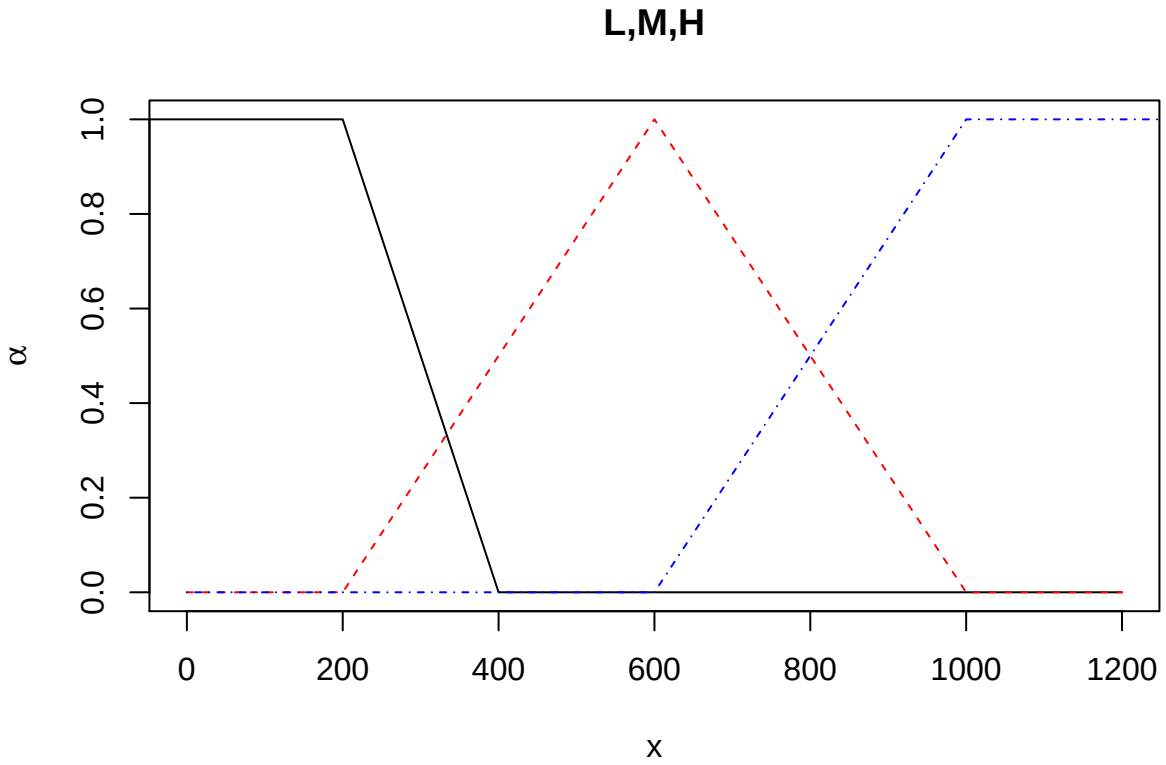
EDUCATION

```
L <- TrapezoidalFuzzyNumber(-10^20,8,8,12)
M <- TrapezoidalFuzzyNumber(8,12,12,16)
H<-TrapezoidalFuzzyNumber(12,16,16,10^20)
plot(L,xlim=c(5,20),main="Education L,M,H")
plot(M, add=TRUE, col=2, lty=2)
plot(H, add=TRUE, col=4, lty=4)
```



Premium

```
L <- TrapezoidalFuzzyNumber(-10^20,0,200,400)
M <- TrapezoidalFuzzyNumber(200,600,600,1000)
H<-TrapezoidalFuzzyNumber(600,1000,1000,10^20)
plot(L,xlim=c(0,1200),main="L,M,H")
plot(M, add=TRUE, col=2, lty=2)
plot(H, add=TRUE, col=4, lty=4)
```



The actuaries came up with the inferences rules table below for the premium output based on AGE input and EDUCATION input

	AGE		
EDUCATION	Y	M	A
L	H	H	M
M	M	M	L
H	M	L	L

Supposed a new driver comes in with an age of 24 and an eductation level of Associate degree (14 years)!

The premium is computed as below

A driver :age 24 and education of Associate degree,AG=24, ED= 14

Step 1

If we plot the graphs of memberships and draw perpendicular and horizontal lines, we get AG'=0.5/Y +0.5/M as a fuzzy representation of 24 ED'=0.5/M +0.5 /H as fuzzy representation of 14

step2

Since we will use only Y and M from age and M and H from EducationThe inference table above is simplified into M,M M and L for Yp.Using the max and min rules of defuzzification we get the intermediate value of

Y_p' as

$$Y_p' = 0.5/M + 0.5/L$$

Step3

We are interested in transforming the fuzzy term above into fuzzy numerical term ;To achieve that we draw 2 horizontal lines splitting M and L by 0.5. We draw vertical lines from each value of premiums (200,400,....1200) and we write all coefficient encounter

we get :*

$Yp' = 0/200 + 0.5/400 + 0.5/600 + 0.5/800 + 0.5/1000 + 0.5/1200$

Step4

COG

```
library(knitr)

Yp= round((0.5*400+0.5*600+0.5*800+0.5*1000+0.5*1200)/(0.5+0.5+0.5+0.5+0.5))
Premium=Yp
kable(paste("premium to be paid $",Premium))
```

premium to be paid $ 800

An adult with education of 20 years and AGE of 40

Step1

If we plot the graphs of memberships and draw perpendicular and horizontal lines, we get

$AG' = 1/A$ ***as a fuzzy representation of 40***

$ED' = 1/H$ ***as fuzzy representation of 20***

Step2

Since we will use only A from age and H from Education.The inference table above is simplified into L for Yp.Using the max-min rules of defuzzification we get the intermediate value of

Y_p' ***below***

$$Y_p' = 1/L$$

Step3

We are interested in transforming the fuzzy terms above into fuzzy numerical term .To achieve that we draw 2 horizontal lines splitting L by 1 We draw vertical lines from each value of premiums (200,400,....1200) and we write all coefficient encounter

we get :

$Y'_p = 1/200 + 0.0/400 + 0.0/600 + 0.0/800 + 0.0/1000 + 0.0/1200$

Step4

COG(Center of gravity rule)

```
Yp= round((1*200+0.0*400+0.0*600+0.0*800+0.0*1000+0.0*1200)/(1))
Premium=Yp
kable(paste("Premium to be paid is $",Premium))
```

Premium to be paid is $ 200

b) Singleton method

PURCHASE MADE ONLINE IN RELATION TO BROWSING TIME.

An online company after contacting a third party is able to determine the annual income of its customers.They place also some tools that collect the time spent by users on their platform The analytic department believe their is a relationship between time spent on the platform ,the income of the customer and the purchase made!So the system as BT and IN as inputs respectively as Browsing time and Income and PU as Purchase output For simplification all variables are split into S,M,L as small,medium and large

Browsing time is $BT : \{S, M, L\}$

with values :$\{20, 30, 40, 50, 60, ...\}$

a TFN is used here with

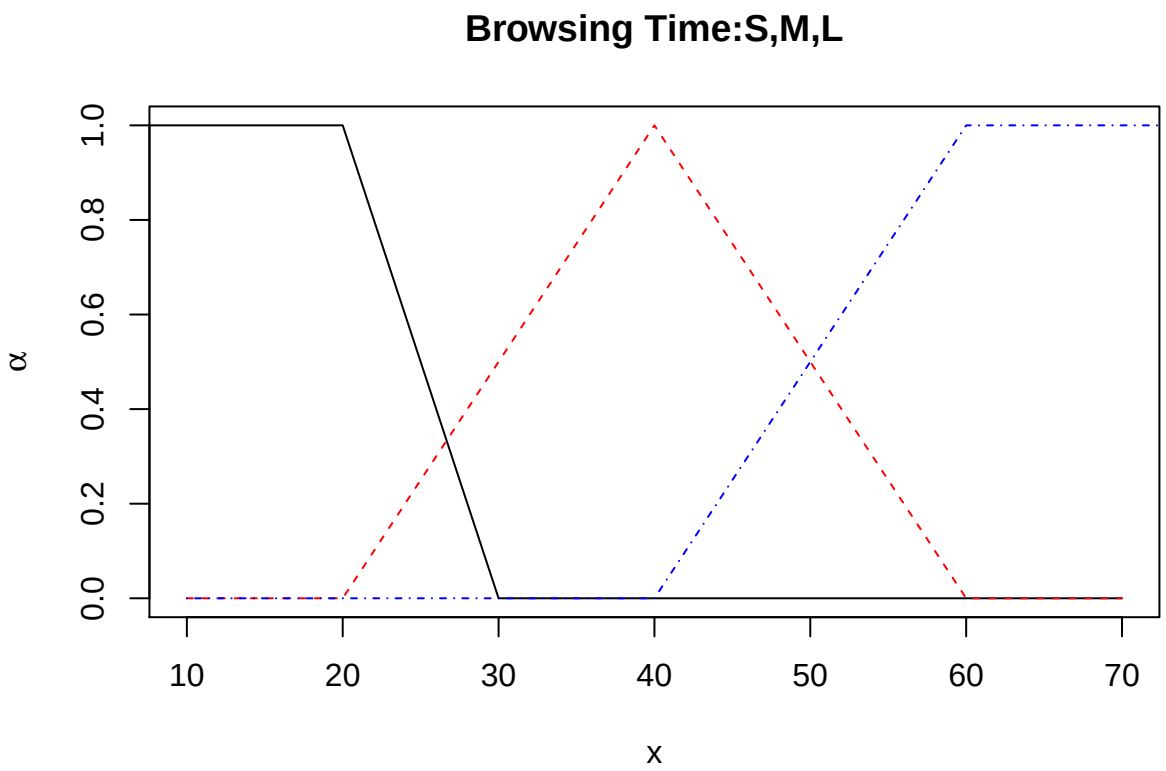
$S : (-\infty, 20, 30)$

$M : (20, 40, 60)$

$L : (40, 60, \infty)$

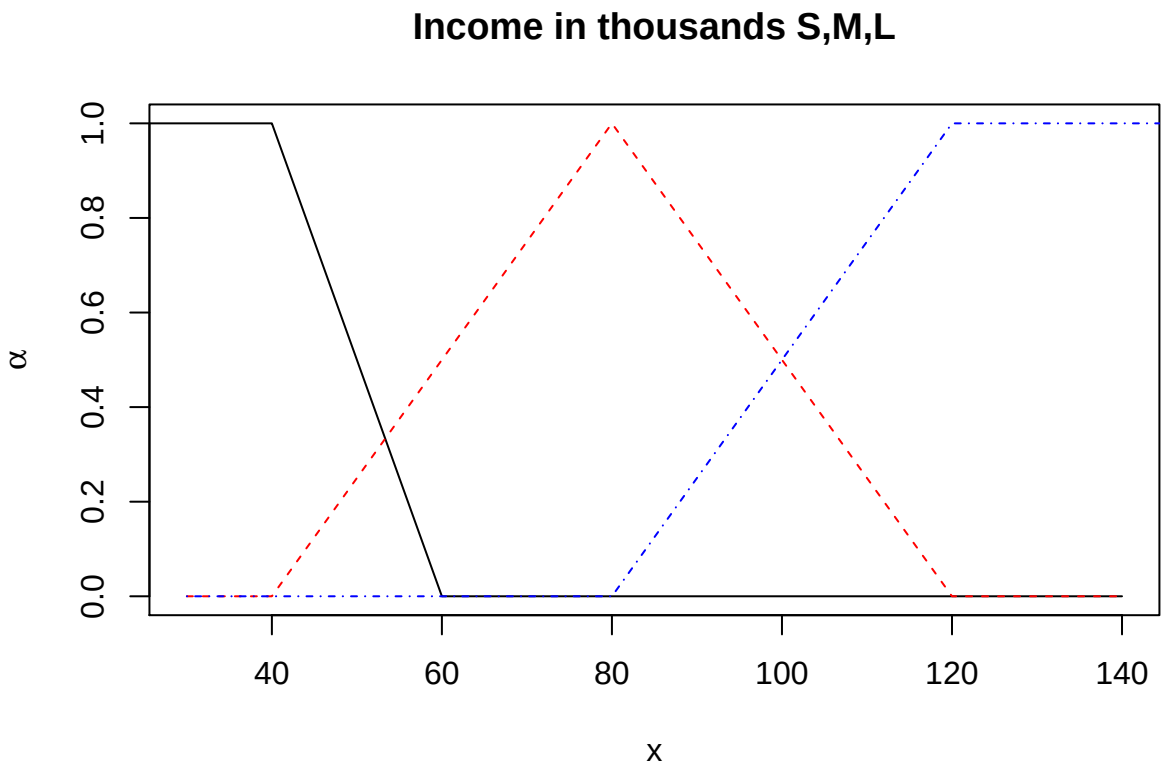
Membership function of Browsong time BT

```
S <- TrapezoidalFuzzyNumber(-10^20,20,20,30)
M <- TrapezoidalFuzzyNumber(20,40,40,60)
L<-TrapezoidalFuzzyNumber(40,60,60,10^20)
plot(S,xlim=c(10,70),main="Browsing Time:S,M,L")
plot(M, add=TRUE, col=2, lty=2)
plot(L, add=TRUE, col=4, lty=4)
```



Membership function of Income

```
S <- TrapezoidalFuzzyNumber(-10^20,40,40,60)
M <- TrapezoidalFuzzyNumber(40,80,80,120)
L<-TrapezoidalFuzzyNumber(80,120,120,10^20)
plot(S,xlim=c(30,140),main="Income in thousands S,M,L")
plot(M, add=TRUE, col=2, lty=2)
plot(L, add=TRUE, col=4, lty=4)
```



Income is IN :$\{S, M, L\}$

with values :$\{40, 60, 80, 100, 120, ...\}$ in thousands

a TFN is used here with

$S : (-\infty, 40, 60)$

$M : (40, 80, 120)$

$L : (80, 120, \infty)$

Purchases rules are summarized in the table below by the Analytics team.The team believe based on past data that purchase values with Income and browsing time are below

INCOME			
BT	S	M	L
S	0	100	200
M	100	300	500
L	200	600	1000

For a customer who spends 50 minutes on the site and with income of 80 thousand, we predict the purchase below:

Step 1

$BT = 50$

$IN = 80$

By drawing vertical and horizontal lines through the values of BT and IN

we get

$BT' = 0.5/M + 0.5/L$

$IN' = 1/M$

Step 2

after simplification and using the max-min rules we get:

$PU' = 0.5/300 + 0.5/600$

Step 4

COG

$PU = round((0.5 * 300 + 0.5 * 600)/0.5 + 0.5)$

```
PU= round((0.5*300+0.5*600)/(0.5+0.5))
Purchase=PU

kable(paste("Predicted Purchase is $",Purchase))
```

Predicted Purchase is $ 450

For a customer who spend 70 minute with an income of 150 thousands below are the calculations

Intuitively one can see from the rules'table that 70 falls into L for browsing time and 150 falls into L for INCOME,so the intersection of L and L gives 1000 for the purchase to be madePU=1000

But let's take the regular route below

Step 1

drawing horizontal and vertical lines give us the values below

$BT' = 1/L$

$IN' = 1/L$

Step2

after crossing out unrelated values we get

$PU" = 1/1000$

Step 4

$$PU = round((1 * 1000)/1)$$

```
PU= round((1*1000)/(1))
Purchase=PU
kable(paste("Purchase will be $ ",Purchase))
```

Purchase will be $ 1000

c)TSK

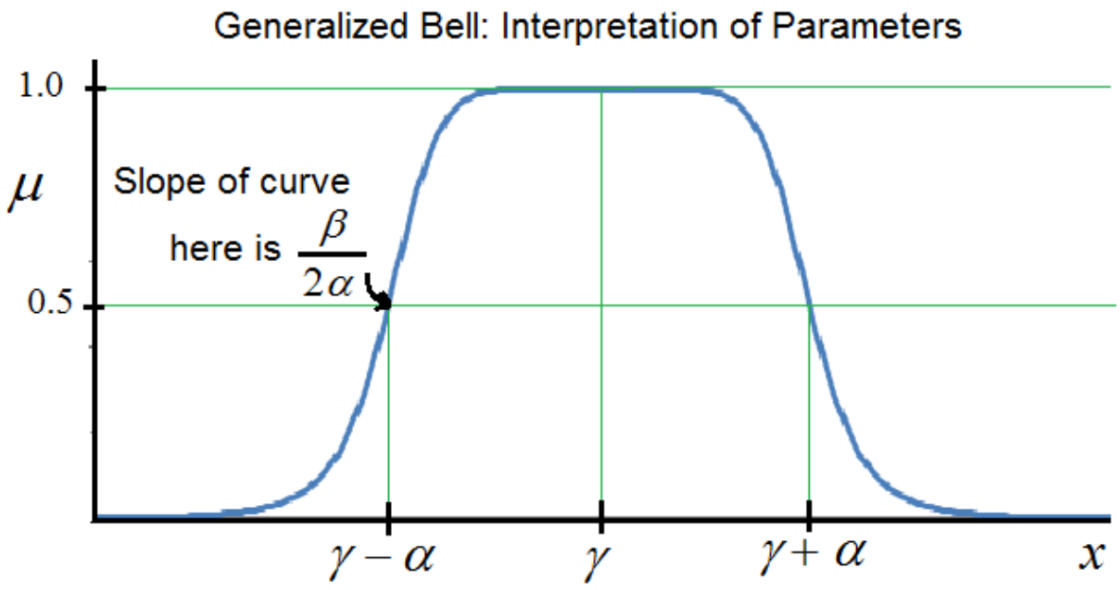
SALES OF PRODUCT BASED ON RATINGS AND BRAND REPUTATION

An online market provider(Amazon for example) is trying to predict the purchase of a categorized product (Suits for example) based on each product ratings, and the brand reputation ratings are between 1 and 5, and brand reputation is between 1 and 5. But somehow ratings are defined by

$x_1 : \{S, M, L\}$

Brand reputation by

$x_2 : \{S, L\}$

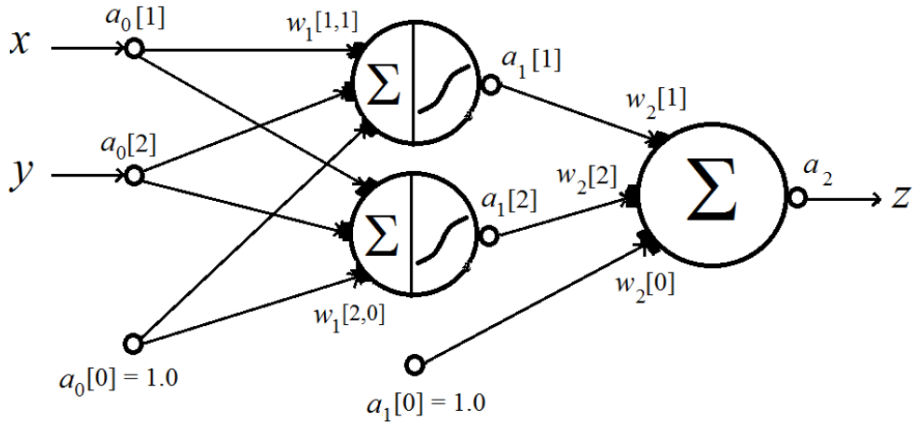


After training,the system parameters are in the tables below

C ij		j	
	0	1	2
1	5	0.5	0.6
2	6	0.4	1
3	2	0.5	0.3
4	10	-0.1	0.8
5	-3	0.8	0.3
6	8	-0.5	1

	α	β	γ
S1	10	1	5
M1	10	0.5	15
L1	8	1	20
S2	10	1.5	5
L2	11	1	18

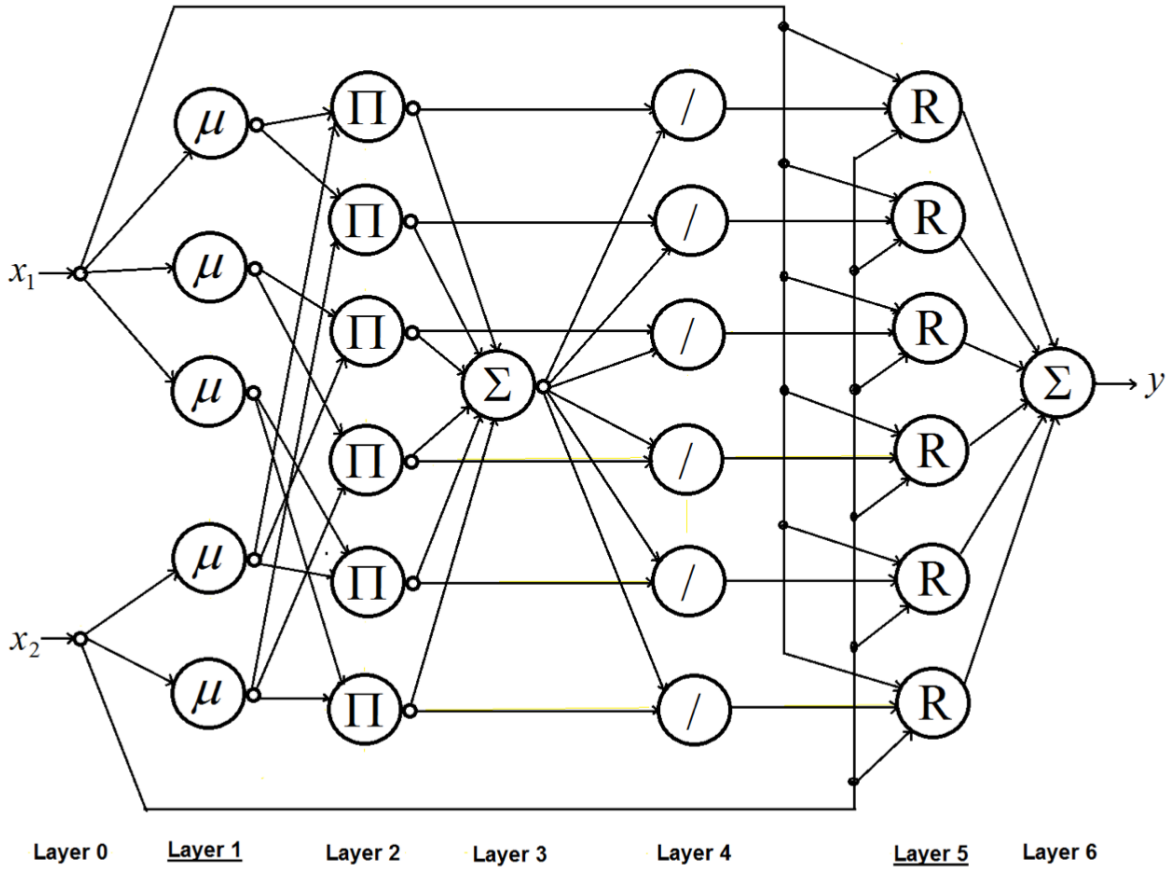
Inputs outputs are related as the schema below



(Note that we are only labeling some of the weights on the diagram. There are actually six hidden layer weights: $w_1[1,0]$, $w_1[1,1]$, $w_1[1,2]$, $w_1[2,0]$, $w_1[2,1]$, and $w_1[2,2]$. There are three output layer weights: $w_2[0]$, $w_2[1]$, and $w_2[2]$.)

ANFIS NETWORK LAYERS BELOW

ANFIS network



m = 2 = number of inputs.

$k[1]$ = 3 = number of linguistic terms used for input x_1(S, M, L).

$k[2]$ = 2 = number of linguistic terms used for input x_2 (S, L).

$n = \prod_{i=1}^{m} k[i]$ = 3·2 = 6 = number of rules.

$a_0[i] = x_i$ for i = 1, . . , m.

$$a_1[i, l] = \frac{1}{1+\left|\frac{a_0[i]-\gamma[i,l]}{\alpha[i,l]}\right|^{2\beta[i,l]}} \quad \text{for } i = 1, \ldots, m,\ l = 1, \ldots, k[i].$$

In this case, we would calculate five a_1 values: $a_1[1, 1]$, $a_1[1, 2]$, $a_1[1, 3]$, $a_1[2, 1]$, and $a_1[2, 2]$. In order to do this, we assume that we have fifteen different parameters available to us based on training: $\alpha[1, 1]$, $\beta[1, 1]$, $\gamma[1, 1]$, $\alpha[1, 2]$, $\beta[1, 2]$, $\gamma[1, 2]$, $\alpha[1, 3]$, $\beta[1, 3]$, $\gamma[1, 3]$, $\alpha[2, 1]$, $\beta[2, 1]$, $\gamma[2, 1]$, $\alpha[2, 2]$, $\beta[2, 2]$, and $\gamma[2, 2]$.

$$a_2[j] = a_1[1, l_1] \cdot a_1[2, l_2] \quad \text{where } j = (l_1 - 1)k[2] + l_2$$
$$\text{for } l_1 = 1, \ldots, k[1] \text{ and } l_2 = 1, \ldots, k[2].$$

This may seem confusing, but what it would mean in this case is simply that

$$a_2[1] = a_1[1, 1] \cdot a_1[2, 1]$$
$$a_2[2] = a_1[1, 1] \cdot a_1[2, 2]$$
$$a_2[3] = a_1[1, 2] \cdot a_1[2, 1]$$
$$a_2[4] = a_1[1, 2] \cdot a_1[2, 2]$$
$$a_2[5] = a_1[1, 3] \cdot a_1[2, 1]$$
$$a_2[6] = a_1[1, 3] \cdot a_1[2, 2].$$

$$a_3 = \sum_{j=1}^{n} a_2[j].$$

$$a_4[j] = \frac{a_2[j]}{a_3} \quad \text{for } j = 1, \ldots, n.$$

$$a_5[j] = a_4[j] \cdot (c[j, 0] + c[j, 1] \cdot a_0[1] + c[j, 2] \cdot a_0[2]) \quad \text{for } j = 1, \ldots, n.$$

This assumes that we have already determined all 18 c values needed from training.

$$a_6 = \sum_{j=1}^{n} a_5[j].$$

Numerical Example: Suppose we have a system with two inputs, x_1 and x_2 , and one input, y, that we have modeled using a TSK-style FIS. Assume the model structure is the following.
Model Structure: $T_{x1} = \{S_1, M_1, L_1\}$ $T_{x2} = \{S_2, L_2\}$ and the membership functions will take the shape of the generalized bell function.
Thus, the number of rules = $|T_{x1}| \cdot |T_{x2}| = 3 \cdot 2 = 6$.

Rule #1: If x_1 is S_1 and x_2 is S_2 then $y = c_{10} + c_{11}x_1 + c_{12}x_2$.
Rule #2: If x_1 is S_1 and x_2 is L_2 then $y = c_{20} + c_{21}x_1 + c_{22}x_2$.
Rule #3: If x_1 is M_1 and x_2 is S_2 then $y = c_{30} + c_{31}x_1 + c_{32}x_2$.
Rule #4: If x_1 is M_1 and x_2 is L_2 then $y = c_{40} + c_{41}x_1 + c_{42}x_2$.
Rule #5: If x_1 is L_1 and x_2 is S_2 then $y = c_{50} + c_{51}x_1 + c_{52}x_2$.
Rule #6: If x_1 is L_1 and x_2 is L_2 then $y = c_{60} + c_{61}x_1 + c_{62}x_2$.

We'll use algebraic product as the fuzzy conjunction operator.

Applying the above into our case we have

Example 1

for a product with rating of 3(x1=3) and well know brand (x2=5) the number of items sold y could be computed as below

x1=3

x2=5

what is y

a0[1]=3

a0[2]=5

```
library(knitr)
a0_1=3
a0_2=5
a1_11=1/(1+abs((a0_1-5)/10)^(2*1))
#
kable(paste("a1[1,1]=",round(a1_11,4)))
```

a1[1,1]= 0.9615

```
a1_12=1/(1+abs((a0_1-15)/10)^(2*0.5))
#a1_12
kable(paste("a1[1,2]=",round(a1_12,4)))
```

a1[1,2]= 0.4545

```
a1_13=1/(1+abs((a0_1-20)/8)^(2*1))
#a1_13
kable(paste("a1[1,3]=",round(a1_13,4)))
```

$$\text{a1[1,3]= 0.1813}$$

```
#So
kable(paste("X1'=",round(a1_11,4),"/S1+",round(a1_12,4),"/M1+",round(a1_13,4),"/L1"))
```

$$\text{X1'= 0.9615 /S1+ 0.4545 /M1+ 0.1813 /L1}$$

```
#
a1_21=1/(1+abs((a0_2-5)/10)^(2*1.5))
#a1_21
kable(paste("a1[2,1]=",round(a1_21,4)))
```

$$\text{a1[2,1]= 1}$$

```
a1_22=1/(1+abs((a0_2-18)/11)^(2*1))
#a1_22
kable(paste("a1[2,2]=",round(a1_22,4)))
```

$$\text{a1[2,2]= 0.4172}$$

```
#So
#X2'
kable(paste("X2'=",round(a1_21,4),"/S2+",round(a1_22,4),"/L2"))
```

$$\text{X2'= 1 /S2+ 0.4172 /L2}$$

```
a2_1=a1_11*a1_21
a2_1
```

```
## [1] 0.9615385
```

```
kable(paste("a2[1]=",round(a2_1,4)))
```

$$\text{a2[1]= 0.9615}$$

```
a2_2=a1_11*a1_22
a2_2
```

```
## [1] 0.4011936
```

```
kable(paste("a2[2]=",round(a2_2,4)))
```

$$\text{a2[2]= 0.4012}$$

```
a2_3=a1_12*a1_21
a2_3
```

```
## [1] 0.4545455
```

```
kable(paste("a2[3]=",round(a2_3,4)))
```

a2[3]= 0.4545

```
a2_4=a1_12*a1_22
#a2_4
kable(paste("a2[4]=",round(a2_4,4)))
```

a2[4]= 0.1897

```
a2_5=a1_13*a1_21
#a2_5
kable(paste("a2[5]=",round(a2_5,4)))
```

a2[5]= 0.1813

```
a2_6=a1_13*a1_22
#a2_6
kable(paste("a2[6]=",round(a2_6,4)))
```

a2[6]= 0.0756

```
a3=a2_1+a2_2+a2_3+a2_4+a2_5+a2_6
#a3
kable(paste("a3=",round(a3,4)))
```

a3= 2.2639

```
a4_1=a2_1/a3
a4_1
```

```
## [1] 0.4247298
```

```
kable(paste("a4[1]=",round(a4_1,4)))
```

a4[1]= 0.4247

```
a4_2=a2_2/a3
kable(paste("a4[2]=",round(a4_2,4)))
```

a4[2]= 0.1772

```
#a4_2
a4_3=a2_3/a3
#a4_3
kable(paste("a4[3]=",round(a4_3,4)))
```

a4[3]= 0.2008

```
a4_4=a2_4/a3
#a4_4
kable(paste("a4[4]=",round(a4_4,4)))
```

a4[4]= 0.0838

```
a4_5=a2_5/a3
#a4_5
kable(paste("a4[5]=",round(a4_5,4)))
```

a4[5]= 0.0801

```
a4_6=a2_6/a3


#a4_6
kable(paste("a4[6]=",round(a4_6,4)))
```

a4[6]= 0.0334

```
#Layer 5

a5_1=a4_1*(5+0.50*a0_1+0.6*a0_2)
#a5_1
kable(paste("a5[1]=",round(a5_1,4)))
```

a5[1]= 4.0349

```
a5_2=a4_2*(6+0.40*a0_1+1.0*a0_2)
#a5_2
kable(paste("a5[2]=",round(a5_2,4)))
```

a5[2]= 2.162

```
a5_3=a4_3*(2+0.50*a0_1+0.3*a0_2)
#a5_3
kable(paste("a5[3]=",round(a5_3,4)))
```

a5[3]= 1.0039

```
a5_4=a4_4*(10-0.10*a0_1+0.8*a0_2)
#a5_4
kable(paste("a5[4]=",round(a5_4,4)))
```

a5[4]= 1.1477

```
a5_5=a4_5*(-3+0.8*a0_1+0.3*a0_2)
#a5_5
kable(paste("a5[5]=",round(a5_5,4)))
```

a5[5]= 0.0721

```
a5_6=a4_6*(8-0.50*a0_1+1.00*a0_2)
#a5_6
kable(paste("a5[6]=",round(a5_6,4)))
```

a5[6]= 0.3843

```
#Layer 6
y=a5_1+a5_2+a5_3+a5_4+a5_5+a5_6
#y
kable(paste("Y,The number of items sold is =",round(y)))
```

Y,The number of items sold is = 9

Example 2

For an excellent rating Suit(5) and a reputable brand , we have the calculations below

x1=5

x2=5

what is y

a0[1]=5

a0[2]=5

```
library(knitr)
a0_1=5
a0_2=5
a1_11=1/(1+abs((a0_1-5)/10)^(2*1))
#
kable(paste("a1[1,1]=",round(a1_11,4)))
```

a1[1,1]= 1

```
a1_12=1/(1+abs((a0_1-15)/10)^(2*0.5))
#a1_12
kable(paste("a1[1,2]=",round(a1_12,4)))
```

$$\text{a1[1,2]= 0.5}$$

```
a1_13=1/(1+abs((a0_1-20)/8)^(2*1))
#a1_13
kable(paste("a1[1,3]=",round(a1_13,4)))
```

$$\text{a1[1,3]= 0.2215}$$

```
#So
kable(paste("X1'=",round(a1_11,4),"/S1+",round(a1_12,4),"/M1+",round(a1_13,4),"/L1"))
```

$$\text{X1'= 1 /S1+ 0.5 /M1+ 0.2215 /L1}$$

```
#
a1_21=1/(1+abs((a0_2-5)/10)^(2*1.5))
#a1_21
kable(paste("a1[2,1]=",round(a1_21,4)))
```

$$\text{a1[2,1]= 1}$$

```
a1_22=1/(1+abs((a0_2-18)/11)^(2*1))
#a1_22
kable(paste("a1[2,2]=",round(a1_22,4)))
```

$$\text{a1[2,2]= 0.4172}$$

```
#So
#X2'
kable(paste("X2'=",round(a1_21,4),"/S2+",round(a1_22,4),"/L2"))
```

$$\text{X2'= 1 /S2+ 0.4172 /L2}$$

```
a2_1=a1_11*a1_21
a2_1
```

```
## [1] 1
```

```
kable(paste("a2[1]=",round(a2_1,4)))
```

$$\text{a2[1]= 1}$$

```
a2_2=a1_11*a1_22
a2_2
```

```
## [1] 0.4172414
```

```
kable(paste("a2[2]=",round(a2_2,4)))
```

a2[2]= 0.4172

```
a2_3=a1_12*a1_21
a2_3
```

```
## [1] 0.5
```

```
kable(paste("a2[3]=",round(a2_3,4)))
```

a2[3]= 0.5

```
a2_4=a1_12*a1_22
#a2_4
kable(paste("a2[4]=",round(a2_4,4)))
```

a2[4]= 0.2086

```
a2_5=a1_13*a1_21
#a2_5
kable(paste("a2[5]=",round(a2_5,4)))
```

a2[5]= 0.2215

```
a2_6=a1_13*a1_22
#a2_6
kable(paste("a2[6]=",round(a2_6,4)))
```

a2[6]= 0.0924

```
a3=a2_1+a2_2+a2_3+a2_4+a2_5+a2_6
#a3
kable(paste("a3=",round(a3,4)))
```

a3= 2.4397

```
a4_1=a2_1/a3
a4_1
```

```
## [1] 0.409884
```

```
kable(paste("a4[1]=",round(a4_1,4)))
```

a4[1]= 0.4099

```
a4_2=a2_2/a3
kable(paste("a4[2]=",round(a4_2,4)))
```

a4[2]= 0.171

```
#a4_2
a4_3=a2_3/a3
#a4_3
kable(paste("a4[3]=",round(a4_3,4)))
```

a4[3]= 0.2049

```
a4_4=a2_4/a3
#a4_4
kable(paste("a4[4]=",round(a4_4,4)))
```

a4[4]= 0.0855

```
a4_5=a2_5/a3
#a4_5
kable(paste("a4[5]=",round(a4_5,4)))
```

a4[5]= 0.0908

```
a4_6=a2_6/a3


#a4_6
kable(paste("a4[6]=",round(a4_6,4)))
```

a4[6]= 0.0379

```
#Layer 5

a5_1=a4_1*(5+0.50*a0_1+0.6*a0_2)
#a5_1
kable(paste("a5[1]=",round(a5_1,4)))
```

a5[1]= 4.3038

```
a5_2=a4_2*(6+0.40*a0_1+1.0*a0_2)
#a5_2
kable(paste("a5[2]=",round(a5_2,4)))
```

a5[2]= 2.2233

```
a5_3=a4_3*(2+0.50*a0_1+0.3*a0_2)
#a5_3
kable(paste("a5[3]=",round(a5_3,4)))
```

a5[3]= 1.2297

```
a5_4=a4_4*(10-0.10*a0_1+0.8*a0_2)
#a5_4
kable(paste("a5[4]=",round(a5_4,4)))
```

a5[4]= 1.1544

```
a5_5=a4_5*(-3+0.8*a0_1+0.3*a0_2)
#a5_5
kable(paste("a5[5]=",round(a5_5,4)))
```

a5[5]= 0.2269

```
a5_6=a4_6*(8-0.50*a0_1+1.00*a0_2)
#a5_6
kable(paste("a5[6]=",round(a5_6,4)))
```

a5[6]= 0.3977

```
#Layer 6
y=a5_1+a5_2+a5_3+a5_4+a5_5+a5_6
#y
kable(paste("Y,The number of items sold is =",round(y)))
```

Y,The number of items sold is = 10

d)

Sometimes the system we are study has variables with characteristics that somehow are related quadratically,a Takagi-Sugeno Fuzzy model whose consequences include second order terms could be used to approximate the non-linear behavior present in that system!

e)

The examples I gave above represent my current research interests.This endenscore how important the approaches taken here could serve me as great tools for developping algorithms and tools for predictions. All the above fuzzy models are useful and applicable in my research.

L-2

a)

When facing a random process or situation, there are many ways to model it using probability repartitions (probability distributions). The probability distribution that gives no assumptions outside of the prior knowledge of the situation returns the maximum entropy. The most ignorant repartition of the probabilities beyond previous data reflecting the case provides the maximum entropy, in other words, the distribution that makes the least claim of knowledge about the situation.

b)

Constraints on mass

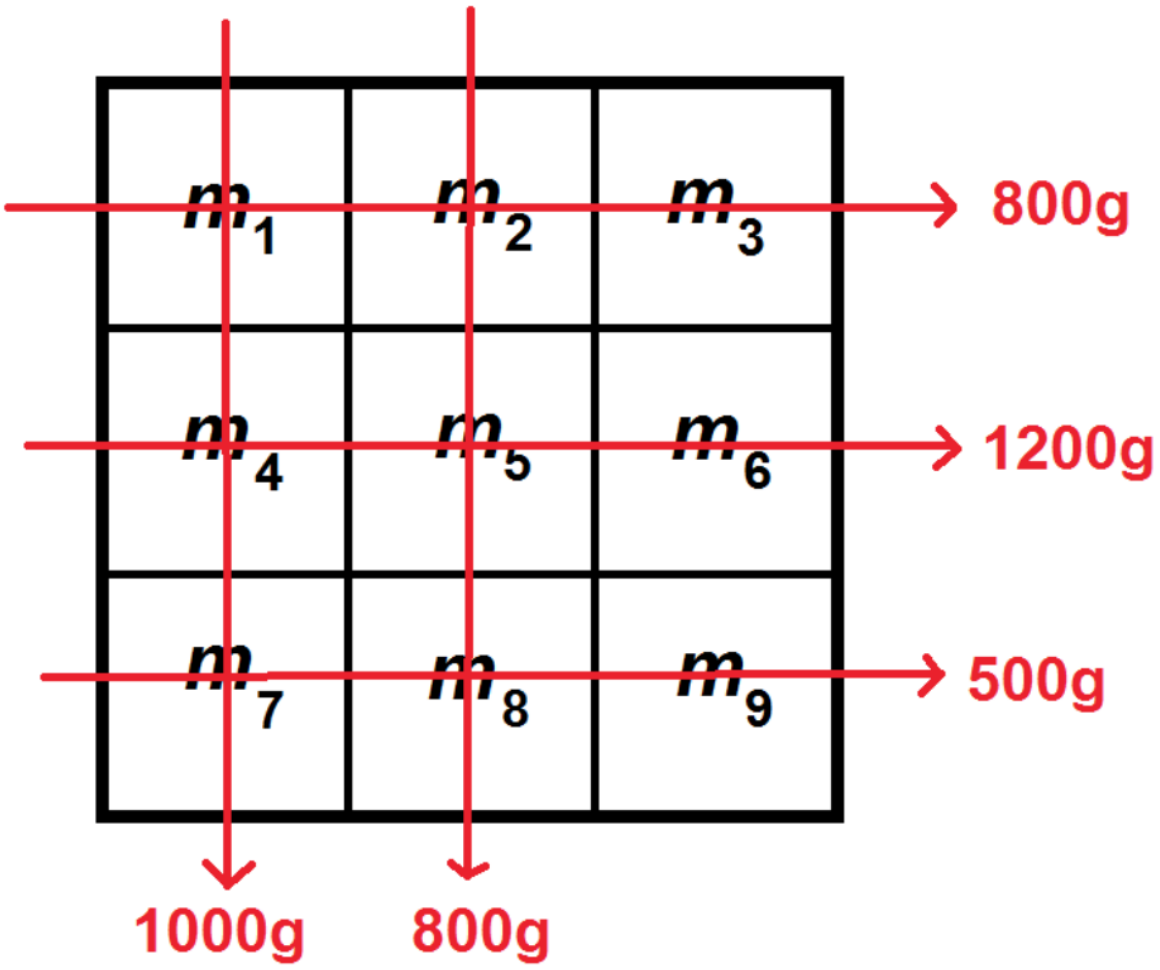


Diagram of cross section for part b of Question L-2

The entropy is defined as

$$H(p) = -\sum_i p_i log(p_i)$$

Subject to $p_i \geq 0$

$\sum_i (p_i) = 1$

$\sum_i p_i r_{ij} = \alpha_j$ for $1 \leq j \leq m$

From Lagrangian

$$J(p) = -\sum_i p_i log(p_i) + \lambda_0(\sum_i p_i - 1) + \sum_j \lambda_j(\sum_i p_i r_{ij} - \alpha_j)$$

I take the derivative with respect to p_{i}:

$$-1 - logp_i + \lambda_0 + \sum_{j=1} \lambda_j r_{ij}$$

I set this to zero and the solution is maximum entropy distribution is

$$p_i* = \frac{e^{\sum_{j=1} \lambda_j r_{ij}}}{e^{1-\lambda_0}}$$

λ_0, λ_1...such that $\sum_i p_i = 1$ and

$\sum_i p_i * r_{ij} = \alpha_j$

In our case here

I define

$p_i = m_i/2500$ being the proportion of mass present in the i section.

So $\sum_i p_i$=1

To get r_{ij}, I use the constraints given

$p_1 + p_2 + p_3 = 800/2500$

$p_4 + p_5 + p_6 = 1200/2500$

$p_7 + p_8 + p_9 = 500/2500$

$p_1 + p_4 + p_7 = 1000/2500$

$p_2 + p_5 + p_8 = 800/2500$

The above equations didn't yield any clear solution.So I decide to use trial and error method to solve the problem.I tried to make the least assumption as possible so the entropy could be maximum The way the problem is formulated with constraints if we are to assign masses to each block and we start with we can decide the value of m1 and m2 to meet the constraint we cannot decide of the value of m3, also m4 and m5 could be randomly chosen but other values can't.Taking the constraints into consideration, we end up with the equations

$m_3 = 800 - m_1 - m_2$

$m_6 = 1200 - m_4 - m_5$

$m_7 = 1000 - m_1 - m_4$

$m_8 = 800 - m_2 - m_5$

$m_9 = 700 - m_6 - m_3$

So I designed a way to start with m1,m2 then m4 m5, and I computed the entropy of the system at each step incremental or decremental I sometimes go by step of 10,20 or 1 depending on the value of the entropy returned.I reiterate the process until I get a maximum entropy of:

3.066906

Then I recorded all the values of the masses

$m_1 = 329$

$m_2 = 258$

$m_3 = 213$

$m_4 = 465$

$m_5 = 381$

$m_6 = 354$

$m_7 = 206$

$m_8 = 161$

$m_9 = 133$

```
library(knitr)
library(pander)
    #m1=((800)/3+(1000/3))/2 I used this initial starting value
m1=329### 329 Good entropy!! "3.066906"
    #m2=((800)/3+(800/3))/2 I used this initial value
m2= 258# 258The best
m3=(800-m1-m2)
m4=465### 465# Very great Good!!
    #m5=(1200-m4)/2
m5=381#### 381 the best value
m6=(1200-m4-m5)
m7=1000-m1-m4
m8=800-m2-m5
m9=700-m6-m3
mass=c(m1,m2,m3,m4,m5,m6,m7,m8,m9)
#mass
mt=matrix(c(m1,m2,m3,m4,m5,m6,m7,m8,m9),byrow=TRUE,nrow=3)
kable(mt,caption="Estimated Masses")
```

Table 78: Estimated Masses

329	258	213
465	381	354
206	161	133

```
pi=mass/2500
pander(caption="Porportion of Masses",pi)
```

0.1316, *0.1032*, *0.0852*, *0.186*, *0.1524*, *0.1416*, *0.0824*, *0.0644* and *0.0532*

```
H=-sum(pi*log2(pi))
kable(paste("THE ENTROPY IS:",round(H,5)))
```

THE ENTROPY IS: 3.06691

According to Klir there are many ways of measuring uncertainty and using maximum entropy is a special case.The Generalized information measure is the sum of Generalized Hartley Measure + the Generalized Shannon measure.

While in the Shannon classical measure we assume that probabilities are adding up to 1,we can define another measure where probabilities are not adding up to 1.If we take a lower bound probability for all elements in the set as baseline ,the sum of those probabilities will be less than 1.If we take the maximum probability and we assigned it to all elements we get a total probability greater than 1.In those case we are talking about unprecised probabilities

In our particular case let's assume that each block cannot be greater than 450 and cannot be less than 150.So the search space is reduce so the maximum entropy but we do meet the conceptual information we want to maximize which is to get values in certain domain

The updated values are

```
m1=390###
m2= 258#

m3=(800-m1-m2)

m4=450###
m5=381####
m6=(1200-m4-m5)
m7=1000-m1-m4
m8=800-m2-m5
m9=700-m6-m3
mass=c(m1,m2,m3,m4,m5,m6,m7,m8,m9)
kable(paste("Minimum mass in this case:",min(mass)))
```

Minimum mass in this case: 152

```
kable(paste("Maximum mass in this case:",max(mass)))
```

Maximum mass in this case: 450

```
#max(mass)
kable(paste("Total mass:",sum(mass)))
```

Total mass: 2500

```
mt=matrix(c(m1,m2,m3,m4,m5,m6,m7,m8,m9),byrow=TRUE,nrow=3)
kable(mt,caption="Estimated New Masses in this case")
```

Table 83: Estimated New Masses in this case

390	258	152
450	381	369
160	161	179

```
#sum(mass)
pi=mass/2500
#kable(caption="Probabilities",pi)
pander(caption="Porportion of Masses",pi)
```

0.156, *0.1032*, *0.0608*, *0.18*, *0.1524*, *0.1476*, *0.064*, *0.0644* and *0.0716*

```
H=-sum(pi*log2(pi))
kable(paste("THE ENTROPY in this case is",round(H,5)))
```

THE ENTROPY in this case is 3.04922

Conclusion.

Uncertainty could be measured by different means, Depending on the function used to do the measure (maximum entropy or Hartley measure) the final results could be different, we can satisfy one requirement and come to a conclusion while using another type of uncertainty measure the results will be express differently

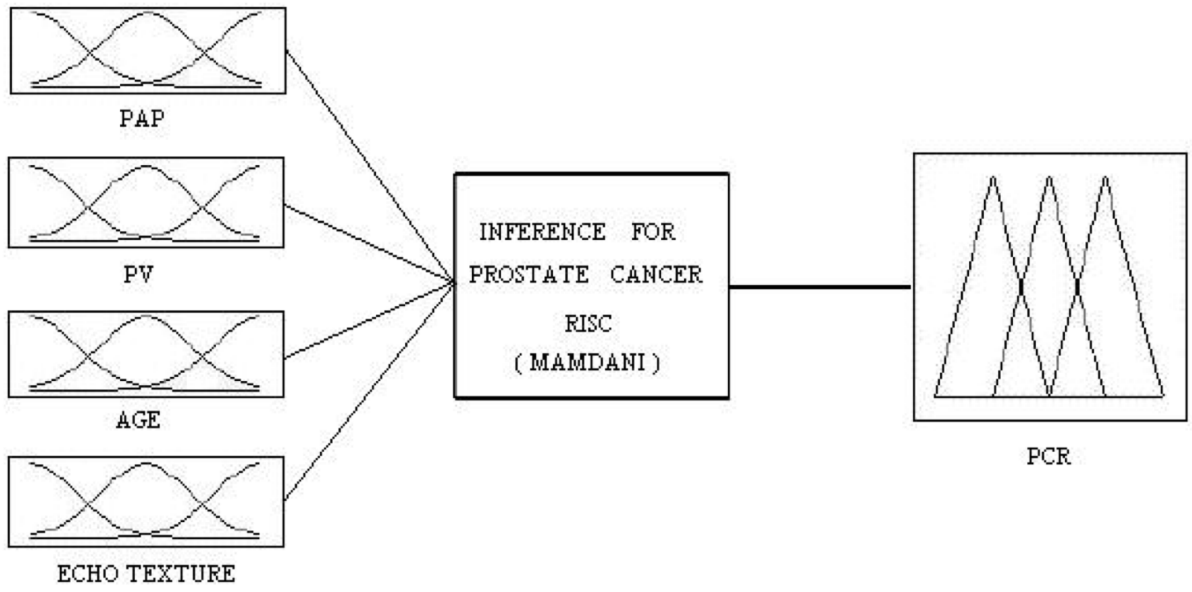
L-3

a)

Hybrid Fuzzy logic and Expert systems used for prostate cancer diagnostic

In artificial intelligence, an expert system is a computer system that emulates the decision-making ability of a human expert. Expert systems are designed to solve complex problems by reasoning through bodies of knowledge, represented mainly as if-then rules rather than through conventional procedural code An expert system is divided into two subsystems: the inference engine and the knowledge base. The knowledge base represents facts and rules. The inference engine applies the rules to the known facts to deduce new facts. Inference engines can also include an explanation and debugging abilities.(Wikipedia) If we build the inference rules part based on fuzzy logic, we have a hybrid system. Medical diagnostics deal with uncertainty, but expert knowledge is crucial, so it is fair to say that we could use a fuzzy expert system which could capture the uncertainty in the outcome better. In this particular example, we will design a fuzzy expert system tool for predicting prostate cancer. we select a pool of individuals with characteristics. Each has four defined parameters those parameters will be the inputs in our system. They are prostate volume, echotexture, total acid phosphate, prostate fraction of acid phosphate).The output which is the predictable outcome is Prostate cancer Risk(PCR), and it is determined using fuzzy expert systems. We use fuzzy logic inferences as done in L1 but this time the rules are guided by an expert doctor. We use this system to write an interactive application or program(previous experts system were written in Prolog), but in our case here I will write a short python code that can be perceived as Fuzzy Expert systems. To limit the length of this report, I will use few samples rules, and I will bypass the Fuzzification and defuzzification parts which are similar to the approach done on part L1

The Structure is below



STRUCTURE OF FUZZY EXPERT SYSTEM

There are many rules in the inference rule table, all done by expert doctor guidance, but for a more straightforward, description below is just a few extracted lines of those rules.

SAMPLE OF RULE GIVEN BY THE DOCTOR

If (ET is N) and (PAP is L) and (TAP is L) and (PV is L) then (PCR is VL)

if (ET is N) and (PAP is VL) and (TAP is VL) and (PV is VL) then (PCR is N)

if (ET is N) and (PAP is VH) and (TAP is VH) and (PV is VH) then (PCR is VH)

let's supposed that we got as inputs

$ET = 0.479$

$PAP = 1.51$

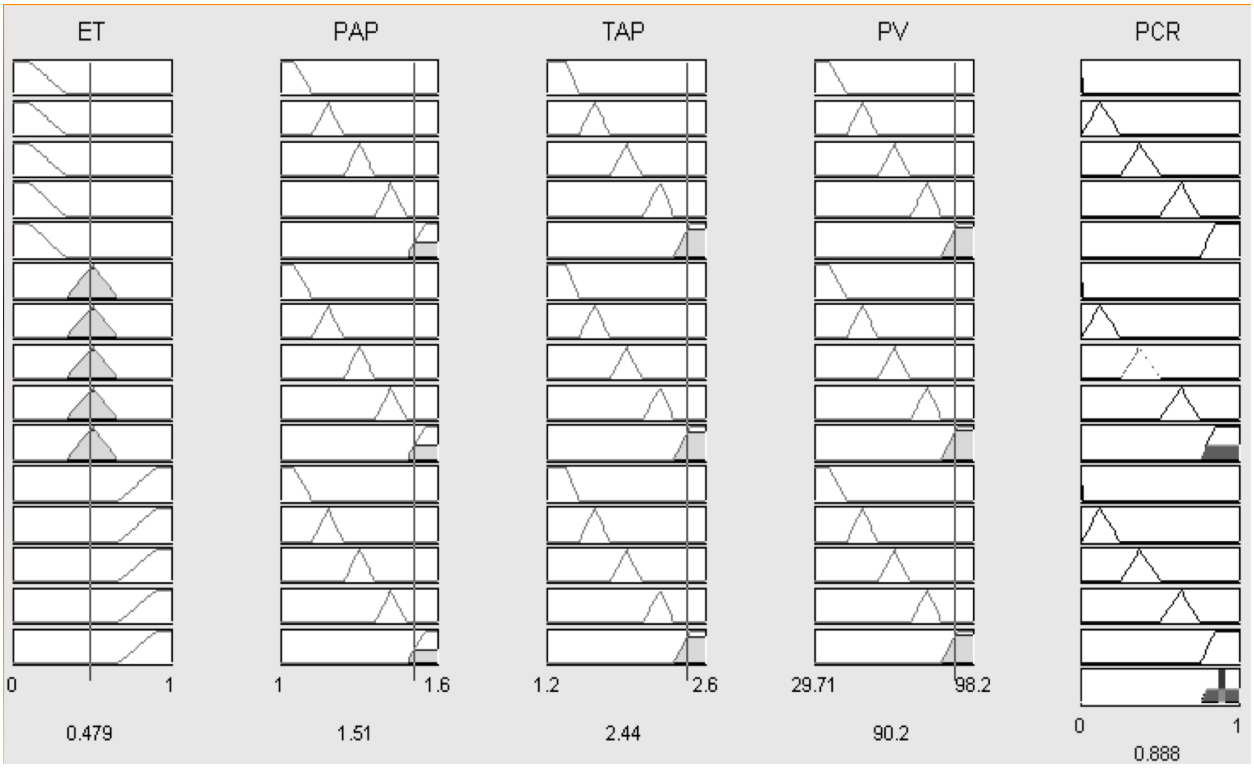
$TAP = 2.44$

$PV = 90.2$

*Through fuzzification defuzzification and center of gravity rules we get **PCR=0.888**

We want the program to be able able to take those 4 inputs and return the PCR.Below is an implementation using python

The rules



This program is just an illustration.Implementing the actual system require all the rules and a function that compute the PCR based on the four inputs using Mandani's style rules of fuzzification and defuzzification. Variables' borders need to be reframed to include the fuzziness of the situation

Sample Implementation of fuzzy expert systems using python, the program is interactive,codes and outputs are below

```
def main ():
    ET_num=eval(input('Enter the value of ET\n'))
    if 0 <ET_num <0.1:
        ET='N'
    elif 0.1<ET_num<0.3:
        ET='L'
    else:
        ET='H'

    PAP_num=eval(input('Enter the value of PAP\n'))
    if 1 <PAP_num <1.2:
        PAP='VL'
    elif 1.2<PAP_num<1.5:
        PAP='L'
    else:
        PAP=PAP_num
        #Suppose  for simplicity that after using the center of gravity rule
        #we get PCR somehow related to PAP by PCR=PAP/1.7
        PCR=PAP/1.7
        PAP='H'
```

```
    TAP_num=eval(input('Enter the value of TAP\n '))
    if  1.2<TAP_num <1.5:
        TAP='VL'
    elif 1.5<TAP_num<2.1:
        TAP='L'
    else:
        TAP='H'

    PV_num=eval(input('Enter the value of PV\n'))
    if 25 <PV_num <50:
        PV='VL'
    elif 50<PV_num<75:
        PV='L'
    else:
        PV='H'


    if ET=='N' and PAP=='L'and TAP=='L' and PV=='L':
        print('The PCR is VL:\n')
    elif ET=='N' and PAP=='VL'and TAP=='VL' and PV=='VL':
        print('The PCR is VL:\n')
    else:
        print('The PCR is:',PCR)
        print('The PCR is  Very High:\n')

if __name__ == '__main__': main()
```

OUTPUT

runfile('/Users/dieudonneouedraogo/ExpertDoctor.py', wdir='/Users/dieudonneouedraogo')

Enter the value of ET

0.479

Enter the value of PAP

1.51

Enter the value of TAP

2.44

Enter the value of PV

90.2

The PCR is: 0.8882352941176471

The PCR is Very High:

b)

Human expertise is essential in many domains as it provides nuances that computers can't offer.When the skills and knowledge of humans are coupled with the power of programming, great applications can be developed.Those fuzzy expert systems could serve as excellent assistants everywhere.A computer app that mimics an expert could be a practical tool and help reduce time and error.For example, an expert system doctor could help provide recommendations to people who don't have access to a real doctor by saving time and avoiding mistake from unexperienced health worker Designing such expert systems could be very cost-effective in the long term, and it can make services more accessible to millions.With my current research interest, this hybrid approach can help me build tools in business, risk management or healthcare by implementing the knowledge in building apps.

References

[1] DR.LEWIS'S COURSE MATERIAL SSIE-519

[2]Ismail SARITAS, Novruz ALLAHVERDI and Ibrahim Unal SERT A Fuzzy Expert System Design for Diagnosis of Prostate Cancer

[3]Wikipidea

PART3 —-DR.DAEHAN WON——

W1.

i)

False

variables can be uncorrelated but dependent

A continuous random variable X has a uniform distribution, denoted U(-1, 1), so its probability density function is:

$$f(x) = \frac{1}{1-(-1)} = \frac{1}{2}; \quad -1 < x < 1$$

Let

$$Y = X^2$$

be another random variable. Clearly X and Y are not independents:

if you know X, you also know Y. And if you know Y, you know the absolute value of X.

X and Y are uncorrelated if the covariance=0

let's compute the covariance.

$$cov(X,Y) = E(XY) - E(X)E(Y)$$

Since the X is uniform between[-1,1] ,its expectation

$$E(X) = \frac{1}{2}(a+b)$$

with a=-1 and b=1 is =0

$$E(X) = 0$$

The covariance is reduced then to

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - 0 * E(Y) = E(X^3) \quad since \quad Y = X^2$$

$$But \quad E(X^3) = 0$$

because the distribution of X is symmetric around zero. Thus the correlation

$$\rho(X,Y) = \frac{Cov(X,Y)}{Sd_X Sd_Y} = 0$$

. We have a situation where the variables are not independent, yet don't have (linear) correlation ($\rho(X,Y) = 0$)

ii)

True

if(Y,X) has a bi-variate normal distribution,X has to be a normal distribution

proof

Let U and V be two independent normal random variables, and consider two new random variables X and Y of the form

$$X = aU + bV$$

$$Y = cU + dV$$

where a, b, c, d, are some scalars. Each one of the random variables X and Y is normal, since it is a linear function of independent normal random variable. Furthermore, because X and Y are linear functions of the same two independent normal random variables, their joint PDF takes a special form, known as the bi-variate normal distribution.

iii)

If the data come from a multivariate normal distribution, then will the Chi Square q-q plot show an upward curve?

False

Not necessarily,it could be upward and ***bended*** showing violation of multinormality

it must be a straight line to maintain multivariate normality

iv)

If a 95% confidence interval is stated as $-1.4 < \mu_1 - \mu_2 < -.2$, then can we state with 95% confidence that $\mu_1$is somewhere between .2 and 1.4 units smaller than μ_2?

True

Proof

here μ_1 represent the mean of population 1 and μ_2 represent the mean of population 2

We know that:

a-If the two confidence intervals do not overlap, we can conclude that there is a statistically significant difference in the two population values at the given level of confidence; or alternatively

b-If the confidence interval for the difference does not contain zero, we can conclude that there is a statistically significant difference in the two population values at the given level of confidence.

Our case fit into b- since the difference is between 2 negative numbers([-1.4,-0.2]).

So We can state with 95% confidence that $\mu_1$is somewhere between .2 and 1.4 units smaller than μ_2(because it is statistically significant)

W2.

i)

We can find α by the likelihood estimation

Below I am changing the variable for easy proof

let

$$\lambda = \frac{1}{\alpha}$$

we have n terms iid sequences X of random variables having exponential distributions

$f_X(x) = \lambda e^{-\lambda x}$

The likelihood function

$$L(\lambda; x_1, ..., x_n) = \lambda^n exp(-\lambda \sum_{j=1}^{n} x_j) \quad ";"in\ here\ means"|"$$

This function help us measure the fitness of the parameter λ for the population distribution.

I derive this from the fact that the variables are iid with PDFs exponential so

$$L(\lambda; x_1, x_2, ...x_n) = \prod_{i=1}^{n} f_X(x_j; \lambda)$$

$$= \prod_{j=1}^{n} \lambda exp(-\lambda x_j)$$

$$L(\lambda; x_1, ..., x_n) = \lambda^n exp(-\lambda \sum_{j=1}^{n} x_j)$$

To find the estimator,I compute the derivative of the log-likelihood (because it is easier than the likelihood) with respect to λ and equal to zero after manual calculation I find the estimator

$$\lambda = \frac{n}{\sum x_j}$$

but since in the beginning I posed

$$\alpha = \frac{1}{\lambda}$$

the estimator

$$\hat{\alpha} = \frac{\sum_{j=1}^{n} x_j}{n}$$

which is the sample mean

ii)

Let the random variable X have the exponential distribution with probability density function $f_X(x) = \frac{1}{\alpha}e^{-x/\alpha}$ with x>0.

The transformation $Y = g(X) = X^{1/\beta}$is a 1-1 transformation from

$X = \{x|x > 0\}$ to $Y = \{y|y > 0\}$ with inverse

$X = g^{-1}(Y) = Y^{1/\beta}$

and Jacobian

$$\frac{dX}{dY} = \beta Y^{\beta-1}.$$

By the transformation rule we have

$$f_Y(y) = f_X(g^-1(y))|\frac{dx}{dy}|$$

$$= \frac{1}{\alpha}\beta y^{\beta-1}e^{-y^\beta/\alpha}$$

$$= \frac{\beta}{\alpha}y^{\beta-1}e^{-y^\beta/\alpha}$$

$$y > 0$$

which is the probability density function of a Weibull random variable

iii)

Maximum Likelihood Estimation (MLE)

Let x1,x2,...xn be a random sample of size n drawn from a population with probability density function f(x,λ) where$\lambda = (\beta, \alpha)$ is an unknown vector of parameters, so that the likelihood function is defined by

$$L = f((\beta, \alpha)) = \prod_i f(x_i, \lambda)$$

(1)

The maximum likelihood of $\lambda = (\beta, \alpha)$, maximizes L or equivalently, the logarithm of L when

$$\frac{dlnL}{d\lambda} = 0.$$

Using the weibull PDFs,its likelihood function is given as :

$$L(x_1, x_2, ...x_n; \beta, \alpha) = \prod_i^n (\frac{\beta}{\alpha})(\frac{x_i}{\alpha})^{\beta-1}exp[-(\frac{x_i}{\alpha})^\beta]$$

$$= (\frac{\beta}{\alpha})(\frac{x_i}{\alpha})^{n\beta-n}\sum_i^n x_i^{(\beta-1)}exp[-\sum_i^n(\frac{x_i}{\alpha}^\beta)$$

Taking the natural logarithm of both sides

$$lnL = nln(\frac{\beta}{\alpha}) + (\beta - 1)\sum_{i}^{n} x_i - ln(\alpha^{\beta-1}) - \sum_{i}^{n}(\frac{x_i}{\alpha})^{\beta}$$

using partial differentiation with rest to β and α in turn and equating to zero,I get the estimating equations as follows

$$\frac{\partial \ln L}{\partial \beta} = \frac{n}{\beta} + \sum_{i}^{n} lnx_i - \frac{1}{\alpha}\sum_{i}^{n} x_i^{\beta} lnx_i = 0 \quad (a)$$

and

$$\frac{\partial}{\partial \alpha} lnL = -\frac{n}{\alpha} + \frac{1}{\alpha^2}\sum_{i=1}^{n} x_i^{\beta} = 0 \quad (b)$$

From b we obtain an estimator of α as

$$\hat{\alpha_{mle}} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{\beta}} \quad (c)$$

By substituting (c) into (a) we get:

$$\frac{1}{\beta} + \frac{1}{n}\sum_{i}^{n} lnx_i - \frac{\sum_{i=1}^{n} x_i^{\beta} lnx}{\sum_{i=1}^{n} x_i^{\beta}} = 0 \quad (d)$$

(d) is solved to obtain the estimate of β by using Newton-Raphson method or other numerical procedure because (d) does not have a closed form solution.

When $\hat{\beta_{mle}}$ is obtained, the value of $\hat{\alpha}$ follows from (c).

Newton-Raphson Method

Let f(x) be a well-behaved function, and let r be a root of the equation f(x)=0. We start with an estimate x_0 of r From x_0, we produce an improved we hope to estimate x_1. From x_1, we produce a new estimate x_2. From x_2, we produce a new estimate x_3. We go on until we are close enough to r or until it becomes clear that we are getting nowhere. The above general style of proceeding is called iterative.

Of the many iterative(or numerical) root-finding procedures, the Newton-Raphson method, with its combination of simplicity and power, is the most widely used:

Let x_0 be a good estimate of r and let $r = x_0 + h$.

Since the true root is r, and $h = r - x_0$,the number h measures how far the estimatex_0 is from the truth.

Since h is small,we can use the linear (tangent line) approximation to conclude that

$$0 = f(r) = f(x_0 + h) \approx f(x_0) + hf^{'}(x_0)$$

, and therefore ,unless $f^{'}(x_0)$ is close to zero it follows that

$$h \approx -\frac{f(x_0)}{f^{'}(x_0)}$$

$$r = x_0 + h \approx x_0 - \frac{f(x_0)}{f'(x_0)}$$

Our new improved estimate x_1 of r is therefore given by

$$x_1 = x_0 - \frac{f(x_0)}{f^{'}(x_0)}$$

The next estimate x_2 is obtained from x_1 in exactly the same way as x_1 was obtained from x_0 :

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

Continue in this way. If x_n is the current estimate, then the next estimate x_{n+1} is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

W3.

i)

Entropy This measure is roughly speaking the logarithm of the number of typical values that the variable(situation) can take

Since we have 13 unknown balls that can be odd each or normal we have 2* 13 cases =26
Entropy of the system

$$H_S = log(2 * 13) = log(26)$$

What is not known through the deviating ball is which one between 13 balls so we have 13 cases

Entropy of the deviating ball

$$H_b = log(13)$$

ii)

ideal measurements we arranged to get log 3 information(equally probables outcomes) and since 3 log 3 is greater than log13. so 3 measurements could work!

iii)

I split the balls into 4,4 and 5 aside and I put the 4 balls each on each scale if there are the same then the 5 remaining have the odd ball

I split into 2 and 3 and check the 3 with 3 normal ones from the correct 8 balls.if the same so the odd ball is between the 2 remaining balls so with 1 more measure I know the odd ball so in total here 3 measurements.

if the 4 each in the beginning are different I take the heavier part and divided into 2 and check if there are different I proceed to determine to odd with my next measurement so here to 3 measurements are enough.

If the 2 balls are equals in the last part,then the first disregard 4 parts had the odd ball(which was actually lighter) I take back those 4 balls and decide them into 2 balls each and do one more step to determine the odd ball.In this case I have done 4 measurements

BUT SINCE I HAVE 2/3 OF A CHANCE OF HEADING TO 3 measurements and 1/3 HEADING TO 4 ,ideally 3 measurements are enough

W4

i)

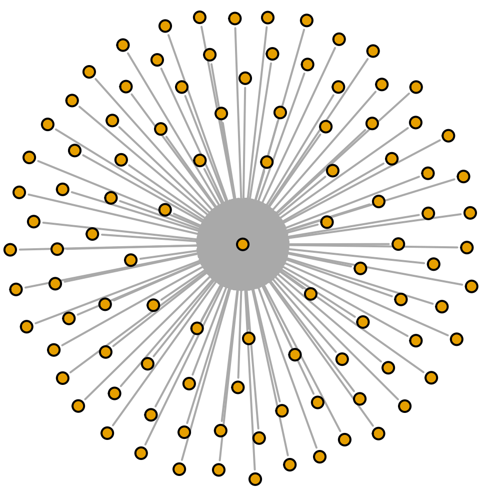
Let N be the number of nodes in the network The central node has **N-1** edges connected to it since it is a star network any other node has 1 edge connected to it so the degree is **1** for those nodes

using R for visual representation we have

igraph is a network science tool for manipulating graphs

Number of nodes =100 so degree distribution is 99 for first node and 1 for other nodes

```
library(pander)
library(knitr) # a nice printing library in R
library(igraph)# here we load the library
net <- make_star(100)# we create a star network with 100 nodes
plot(net, vertex.size=5,vertex.label=NA)#We plot the network with node size 5 without labeling the node
```



```
deg <- degree(net, mode="all") #The degree distribution of all nodes  and we assign it to deg variable
pander(deg,caption="Degree distribution")  #Here  print the degree
```

99, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1* and *1*

ii)

Average degree of the graph

We have N-1 degree for 1 node and N-1 1 degree so the average is

$< k >= \frac{N-1+N-1}{N} = \frac{2N-2}{N}$

So for $N- > \infty < k >= 2$

Using R in our case of 100 nodes we have

We use the mean function in R which give us the average on the variable **deg**

```
Average_degree=mean(deg)
kable(paste("Average degree=",Average_degree))# kable function allows a nicer print out
```

Average degree= 1.98

The result match the formula

$\frac{2*100-2}{100} = \frac{198}{100}$

iii)

For each node: Let n be the number of its neighbor nodes Let m be the number of links among the k neighbors Calculate c = m/(n choose 2) Then C= (the average of c) C indicates the average probability for two of one's friends to be friends too There is no clustering in this type of architecture since we have one central node and N-1 nodes around =>m=0 Clustering coefficient is **zero**

$C = 0$

iv)

In this network we have 2 degrees only **1** and **N-1** k could be chosen between N-1 edges since we have N-1 edges in this type of graph. If we choose k=1, we find surely we find a node with degree 1 along a randomly pick edge if we pick 2,3,....N-2 ,nothing could be found. if we pick N-1 we find 1 node the central node

so

$$q_k = \frac{1+1}{N-1} = \frac{2}{N-1}$$

if $N->\infty$,$q_k = 0$

v)

The Pearson's correlation coefficient of node degrees across links

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

X: degree of start node (in / out)

Y: degree of end node (in / out)

In our case here when we examine every pair of linked nodes we have one node has 1 degree and the other node(central node) has N-1 ,so this is ***a perfect disassortative network*** where very high degree(N-1) is linked to low degree(1) so we have

$$r = -1$$

We can use R to compute it using igraph function: "assortativity_degree()"

This function compute the degree correlation of the network variable we define above ***net***

```
Degree_Correlation=assortativity_degree(net,directed = F)
kable(paste("degree correlation = ",Degree_Correlation))# Degree correlation
```

degree correlation = -1

W5.

Consumer Behavior Analysis In Modern Environment

Abstract

Social networks activities are parts of our lives, and we interact with others through those means more often. As the number of users grows and technology improve, a new reality is taking place: even though most people identify themselves as unique, they tend to show preferences that could be clustered into groups. We are more willing to share our information on the social network; one may argue that we are becoming more predictable than before. We can defined groups or communities by labeling preferences and characteristics as instances where an individual is using perspectives or values as the basis for his or her current behavior. A group could guide the action in some situations.Those groups could be segmented as well.

Within groups either directly or indirectly, studies showed that more experienced members serve as experts and leaders and newer members seek advice and information directly or indirectly as well.We could argue that the collective behavior of the entire network influences any member. Regarding investments being made by companies on social media marketing, it is now indisputable that great insights could be extracted from social networks to drive business decisions and gain competitive advantage. Recent studies in network science reveal the presence of well-defined structures in social networks; an example is the presence of homophily which shows individuals with similarity tending to connect to each other. This paper examines the structural qualities of Social Networks towards the identification of trends in consumers behavior, it gives insights to the characterization of consumer behavior, particularly in the area of predictive analytics.

Keywords:

Consumer Behavior Analysis, Network Science, Data Mining, Trend Discover, Predictive Analytics, Personalization

Literature Review

Twitter has been extensively used on predicting trends in many research because of the relatively small size of its attributes. In a broader sense of social network being used to forecast events, Achrekar et al.[1] used Twitter to predict the trend of the flu virus. They successfully used auto-regression models on tweets to accurately predict the numbers published by the Center for Disease Control (CDC).While the CDC wait to collect actual cases to generate figures, their model could quickly predict outbreak and could be used to save lives.Iyengar et al. were able to predict the start and the end of a set of sports, weather and social activities using SVM classifier and hidden Markov Model on Twitter data.[2] Peng et al. investigated re-tweets patterns using conditional random fields.They defined features as the content influence, network influence and temporal decay factor.The results showed that re-tweet predictions could be substantially improved under social relationships compare to a baseline environment[3]. Gloor, Nann,and Schoder used structural qualities to find betweenness centrality of actors by weighing the context of their positions in the network, they successfully predict long-term trends on the popularity of movies and politicians[4]. Understanding the underline structure of social networks and their relationship to each other is vital on predicting the behavior of nodes, in that sense, Mislove et al. studied the structure of different online social networks. Their results confirm the presence of power-law, small-world, and scale-free properties of online social networks they observe that the in-degree of user nodes tends to match the out-degree[5]. Cantonese et al. analyzed the properties of social networking graphs.They examined scaling laws distribution of friendship and centrality measurements [6]. A useful tool in consumer behavior prediction is “Link Mining”. Algorithms using this technique are designed to support performance among some activities including question answering, information retrieval and web-based data warehousing [7]. Erbs et al. proved that training data and data volume improve performance in link discovery with text-based approaches[8]. Qian et al. used link mining techniques on the Enron mail corpus data and were able to show communities within linked nodes, they were able also to identify ‘common

friends' using cluster analysis.[9]. Other methods explored link-predictions with applications for exploring data, distributed environments, and spam analysis. [10][11][12]. Research in Social Networks is also visible on current search technology including Page Rank and HITS [13][14][15]. Using these techniques, Bharat, Henzinger, and Chakraborti presented variations that utilize web page context to weight pages and links based on relevance.[16][17].Sugiyama et al. used the topological structure of a graph to successfully combine few methods including network, quantitative, semantic, data processing, conversion and visualization-based components [18]. Research in Semantic Web technologies also yielded development in Social Networks. In that sense Zhou, Chen, and Yu combined an ontology-based Social Network along with a statistical learning method towards Semantic Web data using an extended FOAF (friend-of-a-friend) ontology applied as a mediation schema to integrate Social Networks and a hybrid entity reconciliation method to resolve entities of different data sources [19]. Thushar and Thilagam used Semantic Web technology for the identification of associations between multiple domains within a Social Network [20]. Several Relational Learning methods have supported Social Network analysis predicated on the concept of homophily-based associations to support learning. In that context, we have the application of probabilistic modeling [21] collaborative relationship [22] and inference-based approaches [23]. Visualization techniques are being used, and it substantially helps in studying Social Networks dynamics. Batajelj and Mrvar created tools for the visualization of large-scale networks where it is possible to identify vertices and relations between clusters [24]. Noel et al. calculated inter-item distances among combinations of elements from which hierarchical clustering dendrograms are visualized to enhance measurement consistency between clusters and frequent item-sets. They introduced an application of association mining to the visualization of link structures. Important frequently occurring higher-order item-sets are often obscured by the poor pairwise treatment of traditional analysis. The approach they take here involves the discovery of frequently occurring item-sets of arbitrary cardinalities, and the assigning of importance to them according to their support frequencies[25]. Levng et al. created Social Viz which provided users with a means to view frequency relationships among multiple entities in a network [26]. They used frequent pattern mining and visualization techniques. a visualizer called SocialViz is developed for providing users with frequency information on the social relationship among multiple entities in the networks. SocialViz could be used a standalone visualization tool, or as an additional tool to existing visualizers, for social networks exploration[26]. David Alfred et al. used a collection of twitter message to extract metrics that determine the effect of key players and find a correlation between their graph structure and the market share of three primary mobile Operating System[27]. Sharad Goel, and Daniel Goldstein used retail data and applied logistic regression and five-fold cross-validation to compute the likelihood of an individual making a purchase based on his contacts past activities.The results show that individuals with contacts who made a purchase before are more likely to purchase than individuals with connections who did not have any previous purchase[28]. Yoon et al. used S&P companies data from 2010-2015 and math them to 24 million user comments directed at those companies' Facebook posts.They tested hypotheses using fixed effect(FE) and random effects(RE) and dynamic (generalized method of moments) and reached to the conclusion that digital engagement volume has significant positive impact on revenue[29] John et al. cautioned the translation of "liking" on social media into an indicator of an intent to make a purchase.Through their study they discover based on more than 14000 cases, that more features in addition to the button "Like" are needed to make the accurate prediction on the purchase[30]. Chong et al. conducted an experimental study on the consumer engagement behavior(CEB) and were able to show using ordinary least square (OLS) regression models that Facebook and YouTube activities positively correlate with box-office revenue, however, their results are not conclusive for twitter.They proposed and tested a set of metrics[31]. Ding et al. used Pre-released movies "likes" data from Facebook and discovered that more campaign on those Pre-released data increases revenue for a film [32]. Yung et al. propose an experimental model where businesses can target new customers; when a customer visits a store, recommendations are guided using the client social media data. The preliminary results show that companies can generate substantial revenue utilizing this process[33]. Hyunmi et al. investigated the ewom(electronic word-of-mouth) of different social networks using Roger's innovation diffusion model on collected daily data from movies and the results pertain that Twitter influence on box office revenue is more significant in the initial opening stage[34].

Limits of the previous approaches.

Most of the work elaborated above even though spread around different techniques do handle Social Network analysis in a static fashion where nodes or actors dynamics are not taken into account. The arrival or departure time of agents are not taking into consideration, but we believe those features can lead to better insights.The geospatial distribution of the contacts is somehow neglected in the studies.Very often our contacts on networks are spread around the globe and depending on the geographical position, reality and culture can create a barrier that is hard to overcome so we believe a segmented approach can be helpful. Most of the work above also are somehow single network an approach, individual may have preferred using different social networks, and we believe an approach that combines multiple networks analysis may be more conclusive. A handicap with most of those research is the limited scope of Social Networks analysis only.Most of the data available in Social Network are unstructured, a hybrid approach where also structured transaction data are used can lead to better predictions.If a new node joins a network, we can make a recommendation based on his contacts preference, but an old node which has transaction data available in retailer database could have a better recommendation based on both analysis.

Methodology

We defined a framework where nodes are individuals or users. Unlike previous studies, we consider the time factor when a node enter or exit the network. We define and track metrics that define nodes activities: number of comments or "like" and browsing time on social networks related to time.We define similarity measures as characteristics shared by different nodes, and we cluster nodes based on those similarities. We combine the use of multiple networks datasets. We describe a second set of nodes as products, and we specify characteristics that differentiate products. We study the proprieties of nodes or group of nodes as well as their edges(relationships).Unlike previous studies, we consider two situations on multiple Social Networks, a condition where consumer's past transactional data is available and a situation in the absence of transactional data. We propose hybrid models for purchase predictions in either situation. We measure the performance of the models proposed on the actual data.

References

[1] Achrekar, H.; Gandhe, A.; Lazarus, R.; Ssu-Hsin Yu; Benyuan Liu; Predicting Flu Trends using Twitter data Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on Publication Year: 2011 , Page(s): 702 - 707

[2] Peng, Huan-Kai; Zhu, Jiang; Piao, Dongzhen; Yan, Rong; Zhang, Ying; Retweet Modeling Using Conditional Random Fields Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on Publication Year: 2011, pp 336-343

[3] Iyengar, Akshaya; Finin, Tim; Joshi, Anupam; Content-Based Prediction of Temporal Boundaries for Events in Twitter Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom) Publication Year: 2011, pp 186-191

[4] Wasserman, Faust, "Social network analysis: methods and applications" (structural analysis in the social sciences), Cambridge University Press, Cambridge.

[5] Measurement and analysis of online social networks by Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, Bobby Bhattacharjee In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (2007), pp. 29-42,

[6] Cantonese, Salvadore, De Meo, Pasquale, Ferrara, Emilio, Fiumara, Giacomo, Provetti, Alessandro, Crawling Facebook for Social Network Analysis, WIMS'11 May 25-27, 2011 Sogndal Norway

[7] Knowledge Discovery and Retrieval on World Wide Web Using Web Structure Mining

Boddu, Sekhar Babu; Anne, V.P Krishna; Kurra, Rajesekhara Rao; Mishra, Durgesh Kumar; Mathematical/Analytical Modelling and Computer Simulation (AMS), 2010 Fourth Asia International Conference on, 2010, pp: 532-537

[8] Erbs, Nicolai, Zesch, Torsten, Gurevych, Iryna, Link Discovery: A Comprehensive Analysis, 2001 Fifth IEEE International Conference on Semantic Computing

[9] Acar, E.; Dunlavy, D.M.; Kolda, T.G.; Link Prediction on Evolving Data Using Matrix and Tensor Factorizations Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on 2009, pp 262-269

[10]Cai-Rong Yan; Jun-Yi Shen; Qin-Ke Peng; Ding Pan; Parallel Web mining for link prediction in cluster server Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on Volume: 4: 2005 , Page(s): 2291 - 2295 Vol. 4

[11] Caverlee, J.; Webb, S.; Ling Liu; Rouse, W.B.; A Parameterized Approach to Spam-Resilient Link Analysis of the Web Parallel and Distributed Systems, IEEE Transactions on Volume: 20, Issue: 10 2009, pp 1422-1438

[12] Rong Qian; Wei Zhang; Bingni Yang; Detect community structure from the Enron Email Corpus Based on Link Mining, Intelligent Systems Design, and Applications, 2006. ISDA '06. Sixth International Conference on Volume: 2Publication Year: 2006 , Page(s): 850 -855

[13] Web structure mining: an introduction da Costa, M.G., Jr.; Zhiguo Gong; Information Acquisition, 2005 IEEE International Conference on 2005

[14]L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the Web., Technical Report, Stanford Univesity, 1998

[15]NMF: Network Mining Framework Using Topological Structure of Complex Networks Sugiyama, K.; Ohsaki, H.; Imase, M.; Yagi, T.; Murayama, J.; Congress on Services Part II, 2008. SERVICES-2. IEEE Publication Year: 2008, pp 210-211

[16] J. Kleinburg, Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5): 604-632 1999

[17]K. Bharat , M.R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment. In ACM SIGIR International Conference on Research and Development in Information Retrieval, pages 104-111, 1998

[18]S. Chakrabarti, B.Dom, and P.Indyk Enhanced hypertext categorization using hyperlinks. In SIGMOD International Conference on Management of Data pp 307- 318, 1998

[19]Semantic Message Link Based Service Set Mining for Service Composition Anping Zhao; Xiaoyong Wang; Ke Ren; Yuhui Qiu;

Semantics, Knowledge and Grid, 2009. SKG 2009. Fifth International Conference on 2009 , Page(s): 338 - 341

[20]Thushar, A.K.; Thilagam, P.S.; An RDF Approach for Discovering the Relevant Semantic Associations in a Social Network Advanced Computing and Communications, 2008. ADCOM 2008. 16th International Conference on Publication Year: 2008, pp 214- 220

[21] Achim Rettinger Matthias Nickles, Volker Tresp Statistical Relational Learning with Formal Ontologies, ECML PKDD '09 Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II

[22] Kirsten, Mathias, Wrobel, Stefan, Inductive Logic Programming, Lecture Notes in Computer Science, 1998, Volume 1446/1998, 261-270, DOI: 10.1007/BFb0027330

[23] Chunying Zhou; Huajun Chen; Tong Yu; Learning a Probabilistic Semantic Model from Heterogeneous Social Networks for Relationship Identification Tools with Artificial Intelligence, 2008. ICTAI '08. 20th IEEE

International Conference on Volume: 1

[24] Batagelj, Vladimir, Mrvar, Andrej, Pajek: Analysis and visualization of large networks, Graph Drawing Software Book. Junger, P. Mutzel, editors 2003

[25]Noel, S.; Raghavan, V.; Chu, C.-H H.H.; Visualizing association mining results through hierarchical clusters, Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on Publication Year: 2001, pp 425 - 432

[26]Leung, Carson Kai-Sang, Carmichael, Christopher L., Exploring Social Networks: A Frequent-Pattern Visualization Approach, IEEE International Conference on Social Computing, 2010

[27] David Alfred Ostrowski. "Social Network Analysis for Consumer Behavior Prediction". Accessible at: http://worldcomp-proceedings.com/proc/p2012/ICA3445.pdf

[28]Sharad Goel, Daniel C. Goldstein. "Predicting Individual Behavior with Social Networks"

[29] Attracting Comments: Digital Engagement Metrics on Facebook and Financial Performance Gunwoo Yoon, Cong Li, Yi (Grace) Ji, Michael North, Cheng Hong & Jiangmeng Liu Pages 1-14 | Received 28 Apr 2017, Accepted 07 Nov 2017, Published online: 24 Jan 2018 https://doi.org/10.1080/00913367.2017.1405753

[30]Leslie K. John, Oliver Emrich, Sunil Gupta, and Michael I. Norton (2017) Does "Liking" Lead to Loving? The Impact of Joining a Brand's Social Network on Marketing Outcomes. Journal of Marketing Research: February 2017, Vol. 54, No. 1, pp. 144-155. https://doi.org/10.1509/jmr.14.0237

[31]Chong Oh, Yaman Roumani, Joseph K. Nwankpa, Han-Fen Hu Beyond likes and tweets: Consumer engagement behavior and movie box office in social media Information & Management Volume 54, Issue 1, January 2017, Pages 25-37 10.1016/j.im.2016.03.004

[32]Chao Ding,Hsing Kenneth Cheng,Yang Duan,YongJin The power of the "like" button: The impact of social media on box office Decision Support Systems Volume 94, February 2017, Pages 77-84 https://doi.org/10.1016/j.dss.2016.11.002

[33]Yung-Ming Li, Lien-Fa Lin, Chun-Chih Ho A social route recommender mechanism for store shopping support Author links open overlay panel Decision Support Systems Volume 94, February 2017, Pages 97-108

https://doi.org/10.1016/j.dss.2016.11.004

[34]Hyunmi Baek,Sehwan Oh,Hee-DongYang,JoongHo Ahn Electronic word-of-mouth, box office revenue and social media Electronic Commerce Research and Applications Volume 22, March-April 2017, Pages 13-23 https://doi.org/10.1016/j.elerap.2017.02.001