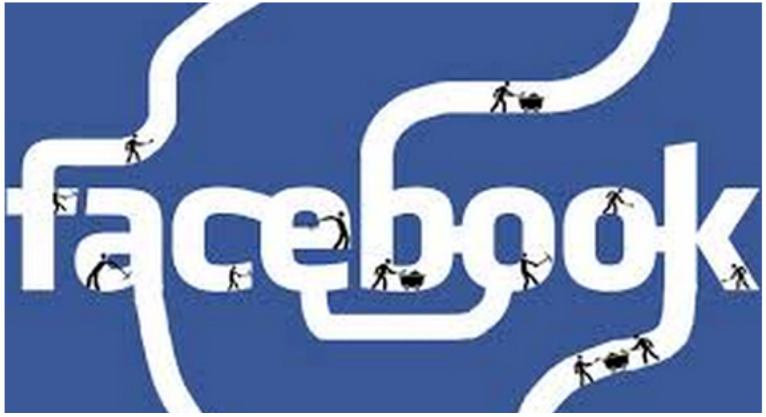
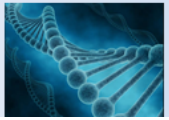
Althea Health Project- Facebook



Presenters: Ram Nagarajan and Kanchan Chauhan
MBA | Data Science & Analytics | Professor: Dr. Sanjiv Das



Agenda

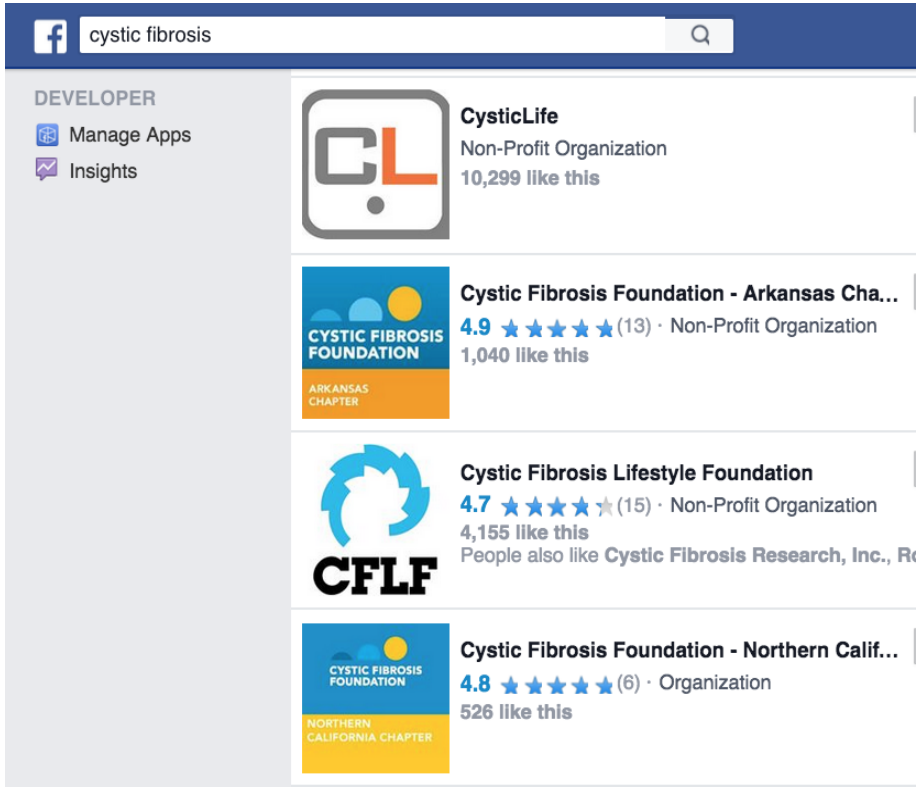
Althea Problem Statement:

"Find Influential Profiles Related to Cystic Fibrosis on Facebook"

- **Althea Overview**
- **Obtain Live Data From FB:**
  - Extract data - public pages & posts
- **Process the extracted data:**
  - Create & plot adjacency matrix
  - Calculate centrality_Ranking
  - Analyse influential profiles
- **Reporting:**
  - Export the files & save the extracted data
  - Plot graphs and charts
- **Future Scope**

Santa Clara University

AltheaHealth & FB Reach for CF



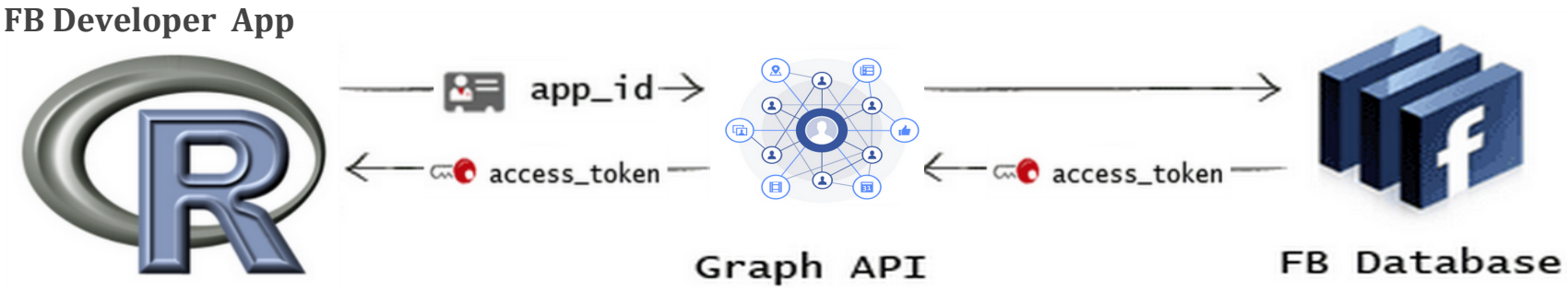
Cystic Fibrosis

- Genetic Rare disease
- Causes production of thick mucus blocking Lungs, Pancreas, Liver, etc
- 30,000 individuals in the U.S. have CF
- 1,000 new cases of CF each year
- $78 million spent in CF research on average each year

ALS Ice Bucket Challenge

Santa Clara University

FB Graph API restriction: effective May 2015



fbOAuth("app_id", "app_secret", extended_permissions = TRUE)

- **FB site search depreciated -> using Rfacebook::searchFacebook()**
- **Demographic data unavailable ->clustering can not be done**
- **Celebrity flag unavailable -> using Rfacebook::getUsers()**
- **Limited data set -> to perform analysis**
- **Multi level info restricted -> dependent on profile's security settings**

Santa Clara University

Extracted Data From Facebook

		Thu_May_28_01/55/10_PM_2015_PublicPages_81391099705.csv							
2	from_id	from_name	message	created_time	type	link	id	likes_count	comments_count	shares_count
3	1.02067E+16	Krista Keller	Anyone also diagnosed with Chr	05-28T18:55:00	status		81391099705	0	0	0
4	2.35861E+11	Angel's Life Foundation - Supporting Cystic Fibrosis Research	Cystic Fibrosis is a genetic illness	05-23T12:24:56	photo	24096064684	23586074609	2	0	0
5	1.51832E+15	Breathe With Me Strawfie Challenge	Veronica has recently had a lung	05-12T22:37:10	video	ube.com/wat	81391099705	1	0	0
6	9.59937E+14	Laura Blaisdell	Hey All! I am doing fundraising f	05-08T11:22:00	status		81391099705	0	0	0
7	1.01533E+16	Michelle Nicol	https://www.justgiving.com/Rac	05-06T18:40:42	link	ng.com/Rach	81391099705	0	0	0
8	81391099705	CysticLife	Hello! I'll be moving to Connecti	05-05T22:07:55	status		81391099705	3	5	0

```
publicPages = getPage(page=pageId,token= tokenKey,n=num,feed=feedFlag)
```

Santa Clara University

Extracted Data From Facebook

```
publicPosts = getPost(pageId, accessToken, n = num, comments = commentFlag, likes = likeFlag)
```

Public Posts = A relational list of 3 lists (Public Posts, Public Comments, Public Likes)

1		PublicPosts_81391099705								
2	from_id	from_name	message	created_time	type	link	id	likes_count	comments_count	shares_count
3	1.5183E+15	Breathe With M	Veronica has	2015-05-12T	video	https://www	8139109970	0	0	0
4	8.4689E+14	Laura Blaisdell	Hey All! I am	2015-05-08T	status		8139109970	0	0	0
5	1.0152E+16	Michelle Nicol	https://www	2015-05-06T	link	https://www	8139109970	0	0	0
6	8.1391E+10	CysticLife	Hello! I'll be	2015-05-05T	status		8139109970	3	5	0

1	PublicLikes	
2	from_name	from_id
3	Heike GÃ¼nther	8.2088E+14
4	Charlie Mannino	1.02E+16
5	Nancy Watson	1.0204E+16
6	Crystal Bello	1.0203E+16

1	Fri_May_28_15/26/18_2015_PublicComments_81391099705.csv					
2	from_id	from_name	message	created_time	likes_count	id
3	1.0202E+16	Karen Carlson	Hello. I'm in Connecticut. Th	15-05-05T22:18:04+00	0	10153274385539706_1
4	1.0205E+16	Elizabeth Gallo	I'm not sure where in Conne	15-05-05T22:36:15+00	0	10153274385539706_1
5	8.1391E+10	CysticLife	Be sure to answer on Cystic	15-05-05T23:09:01+00	0	10153274385539706_1
6	1.0202E+16	Karen Carlson	Thanks. I just did.	15-05-05T23:25:25+00	0	10153274385539706_1

Santa Clara University

Mining FaceBook data - form adjacency matrix

weight = ((#likes * 1) + (#comments * 2) + (celebrity * 100))

Message_Posted_By	Liked_By	Commented_By
John (message 1)	Connie, Sam, Carlina	Carlina
Sam	John, Carlina	Carlina
John (message 2)	Connie	Sam
Connie		John

- **Read celebrity data**

first_nm	last_nm
Miley	Cyrus
Connie	Land
Carlina	Lee
Ram	Nagarajan

From_User	To_User	#Likes	#Comments	Celebrity?	Weight
John	Sam	1			1
John	Connie		1		2
Sam	John	1	1		3
Connie	John	2		1	101
Carlina	John	1	1	1	103

Adjacency Matrix

	John	Sam	Connie	Carlina
John	0	1	2	0
Sam	3	0	0	0
Connie	101	0	0	0
Carlina	103	0	0	0

- Weights used configurable
- Used a threshold percentage to show top influential people
- If provided a list of celebrities, can identify influential people
- Program can be run to mine any public pages/postings

Santa Clara University

Mining FaceBook Data - code snippets

Get page:

```
# Use 81391099705 https://www.facebook.com/pages/CysticLife/81391099705?fref=ts
page = getPage(page="81391099705",token=fb_oauth,n=500,feed=TRUE)
```

Get posts:

```
# Read the post corresponding to the message you are processing
current_post <- getPost(post=page$id[p], n=2000, token=fb_oauth,comments=TRUE,likes=TRUE,n.likes=2000,n.comments=2000)
```

Check for likes and update weights:

```
# Check for existence for a row with same from and to_id. If found, increment the like count, else add
match_index = which((adj_df$to_id == page$from_id[p]) & (adj_df$from_id == current_post$likes$from_id[q]))
# If length(match_index) = 0, it means that no duplicates are found
if (length(match_index) == 0)
{
  if (!is.null(current_post$likes$from_name[q]))
  {
    if (!is.na(current_post$likes$from_name[q]))
    {
     i = i + 1
      adj_df$to_id[i] = page$from_id[p]
      adj_df$to_name[i] = page$from_name[p]
      adj_df$from_id[i]= current_post$likes$from_id[q]
      adj_df$from_name[i] = current_post$likes$from_name[q]
      adj_df$likes[i] = 1
     adj_df$comments[i] = 0
      adj_df$weight[i] = like_weight + adj_df$weight[i]
    }
  }
}
else
  # Else, increment like count for the matching row
{
  adj_df$likes[match_index] = adj_df$likes[match_index] + 1
  adj_df$weight[match_index] = like_weight + adj_df$weight[match_index]
```

Santa Clara University

Mining FaceBook Data - code snippets (contd.)

Celebrity logic:

```
# Execute the following if celebrity check = 1
if (celebrity_check == 1) {
  adj_df_size = i
  for (j in 1:adj_df_size)
  {
    celebrity_match = which(celebrity_df$name == adj_df$from_name[j])
    #print(adj_df$from_name[j])
    if (length(celebrity_match) > 0)
    {
      print(adj_df$weight[j])
      #print(paste(j,(adj_df$from_name[j])))
      adj_df$celebrity[j] = 1
      adj_df$weight[j] = celebrity_weight + adj_df$weight[j]
      print(adj_df$weight[j])
    }
  }
}
```

Apply threshold criteria:

```
# Remove all NA rows
adj_df = adj_df[complete.cases(adj_df),]
# End update

# Now eliminate all rows with weight less than threshold
adj_mod_df = adj_df[adj_df$weight >= weight_threshold,]
# Order the data frame by first name
adj_mod_df = arrange(adj_mod_df,from_name)
```

Unique users:

```
#Capture unique influential users
u_infl_from_ids = unique(adj_mod_df$from_id)# from list users
u_infl_to_ids = unique(adj_mod_df$to_id)# from list users
```

Populate adjacency

```
# Now populate adjacency matrix
size_adj = length(u_infl_df$id)
adj_matrix <- matrix(data=0,nrow=size_adj,ncol=size_adj)
# Now populate adj_matrix
for (a in 1:size)
{
  r = adj_matrx_df$row_indx[a]
  c = adj_matrx_df$col_indx[a]
  adj_matrix[r,c] = adj_matrx_df$weight[a]
}
```

Santa Clara University

Form Adjacency Matrix & Graph

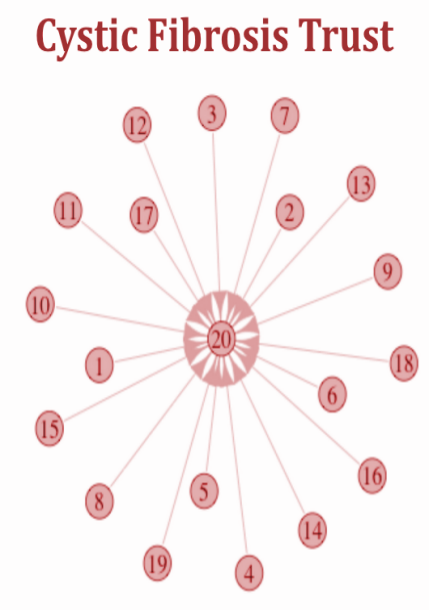
- Form the relationship dataframe with weights
- Eliminate people less than threshold
- Form the list of unique users and then form adjacency matrix/graph
- Plot adjacency graph and relationship graphs
- Determine Centrality_Ranking

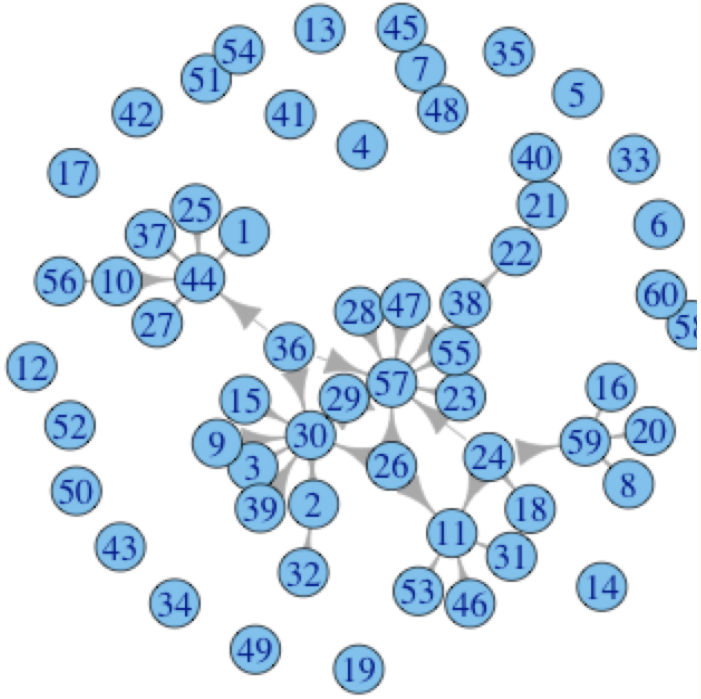
```
g = graph.adjacency(adj_matrix,mode="directed",weighted=TRUE,diag=FALSE)
#Plot the adjacency matrix
plot(g)

# Calculate centrality/page rankingg
centrality_ranking = evcent(g)

# Now add the centrality column to adj_matrix_df
rank_results = u_infl_df
rank_results$ranking = 0
size_cntl = length(rank_results$id)
```

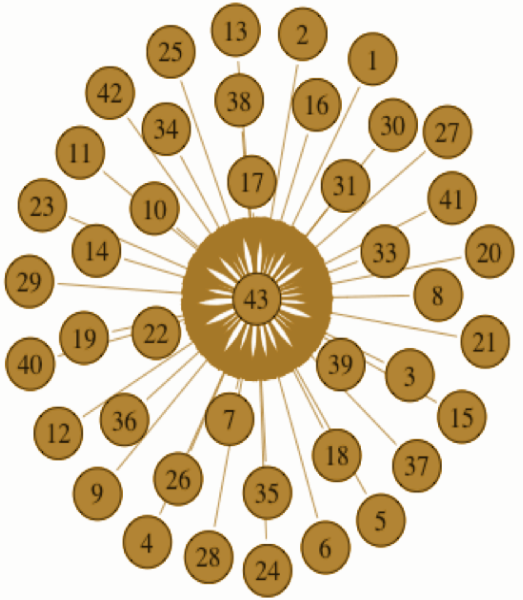
Adjacency Matrix Networks

Cystic Fibrosis Public Group

Cystic Life Org







Santa Clara University

Output Files

Influential_dataset_81391099705							
from_id	from_name	to_id	to_name	likes	comments	celebrity	weight
1.02057E+16	Crystal Bello	8.1391E+10	CysticLife	256	0	0	256
9.79933E+14	Lesa Daisy	8.1391E+10	CysticLife	235	3	0	241
4.12442E+14	Zbigniew Landowski	8.1391E+10	CysticLife	134	0	0	134
1.02005E+16	Charlie Mannino	8.1391E+10	CysticLife	130	0	0	130
1.57242E+15	Ram Nagarajan	1.5183E+15	Breathe With Me Strawfie Challenge	1	0	1	101

weight = ((#likes * 1) + (#comments * 2) + (celebrity * 100))

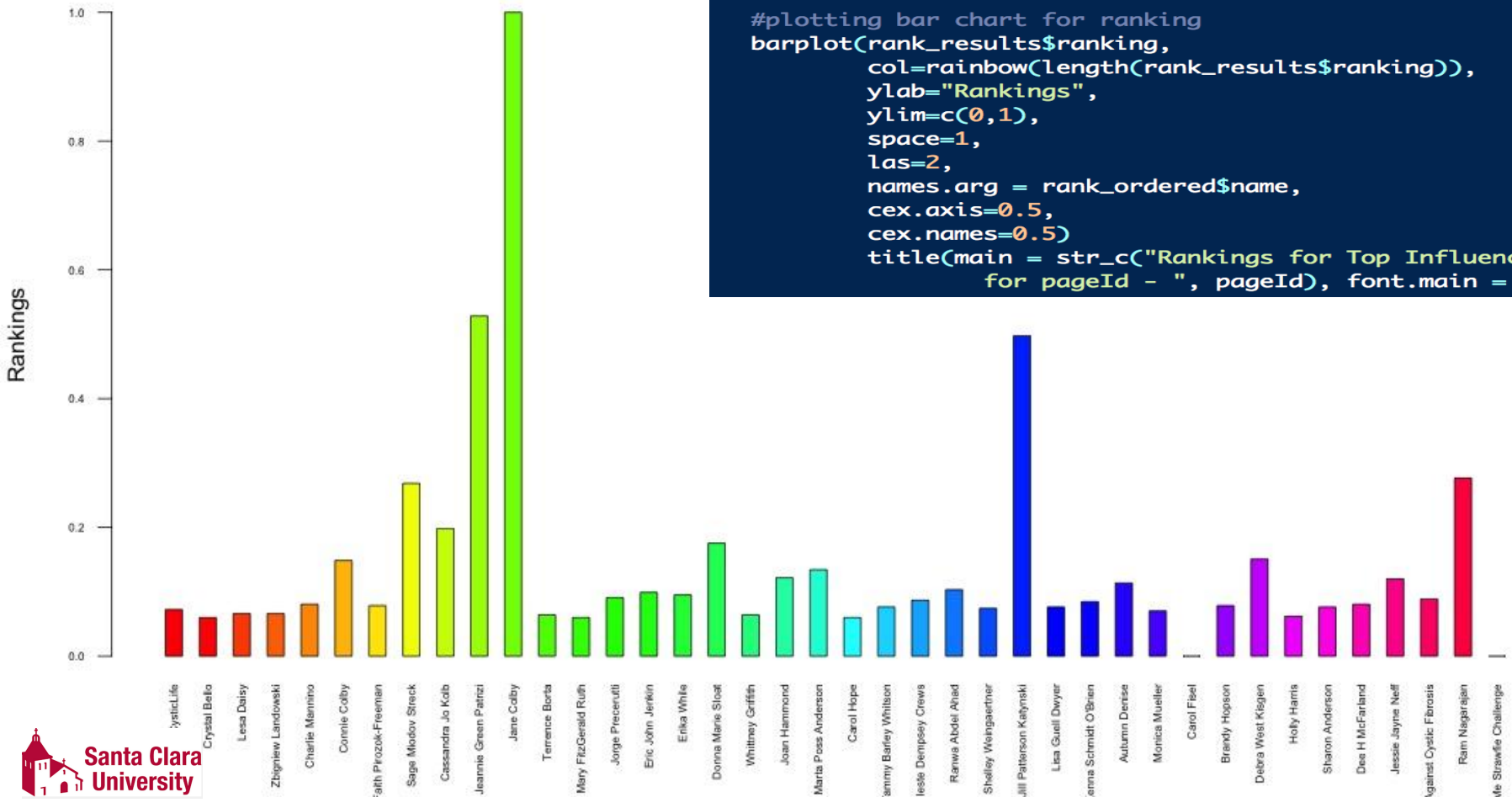
Ranking_81391099705			
id	name	index_subscript	ranking
81391099705	CysticLife	11	1
1.02057E+16	Crystal Bello	10	0.528336167
9.79933E+14	Lesa Daisy	26	0.49737897
4.12442E+14	Zbigniew Landowski	39	0.276550962
1.02005E+16	Charlie Mannino	8	0.26829571
8.05589E+14	Connie Colby	9	0.198126063

- Ranking is a measure of influence of nodes (people) on others
- We use evcent() on adjacency graph to calculate ranking

Santa Clara University

Rankings for Top Influencers for pageId - 81391099705

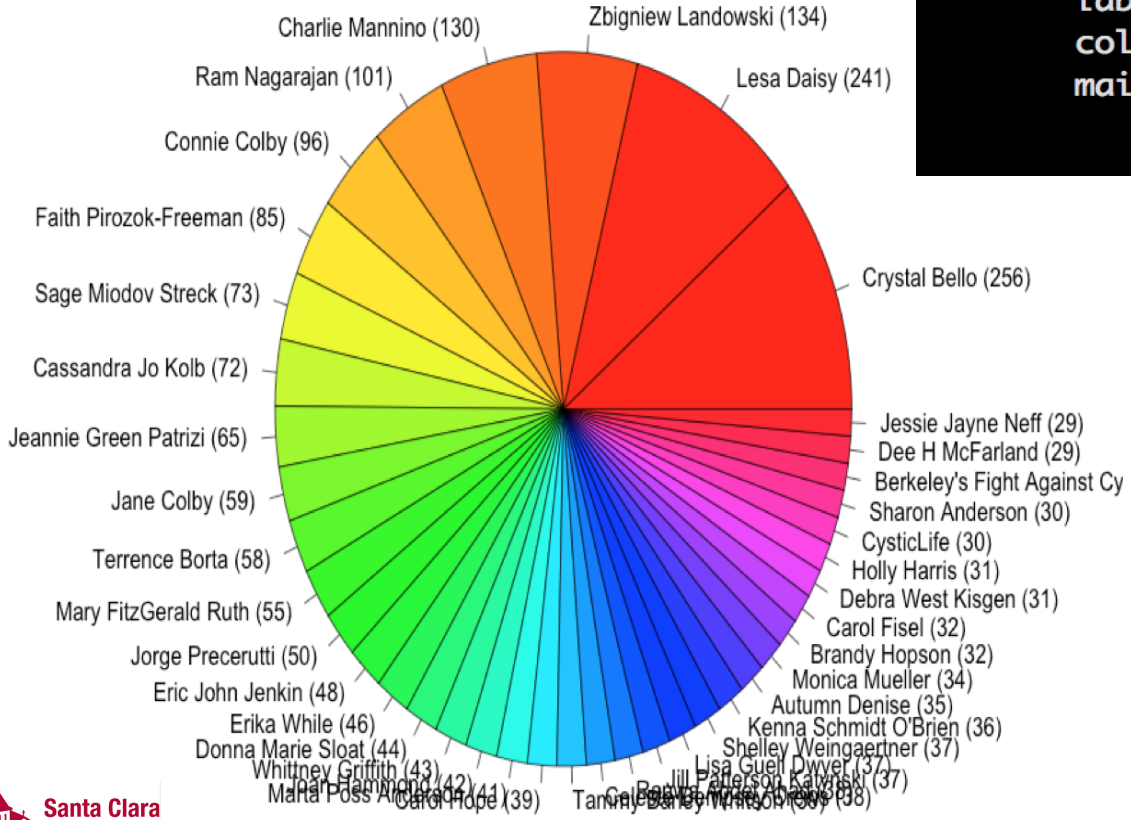
```
#plotting bar chart for ranking
barplot(rank_results$ranking,
        col=rainbow(length(rank_results$ranking)),
        ylab="Rankings",
        ylim=c(0,1),
        space=1,
        las=2,
        names.arg = rank_ordered$name,
        cex.axis=0.5,
        cex.names=0.5)
        title(main = str_c("Rankings for Top Influencers
                for pageId - ", pageId), font.main = 4)
```

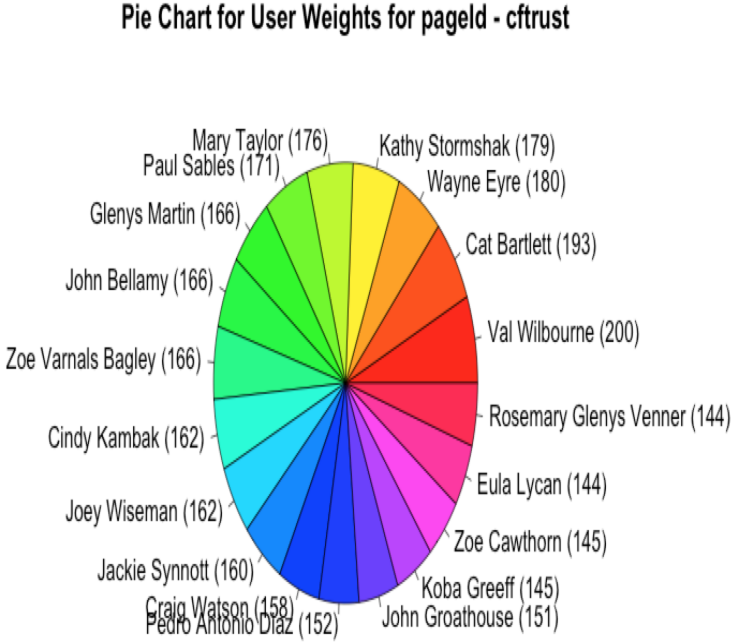


Santa Clara University

Page wise Pie Charts

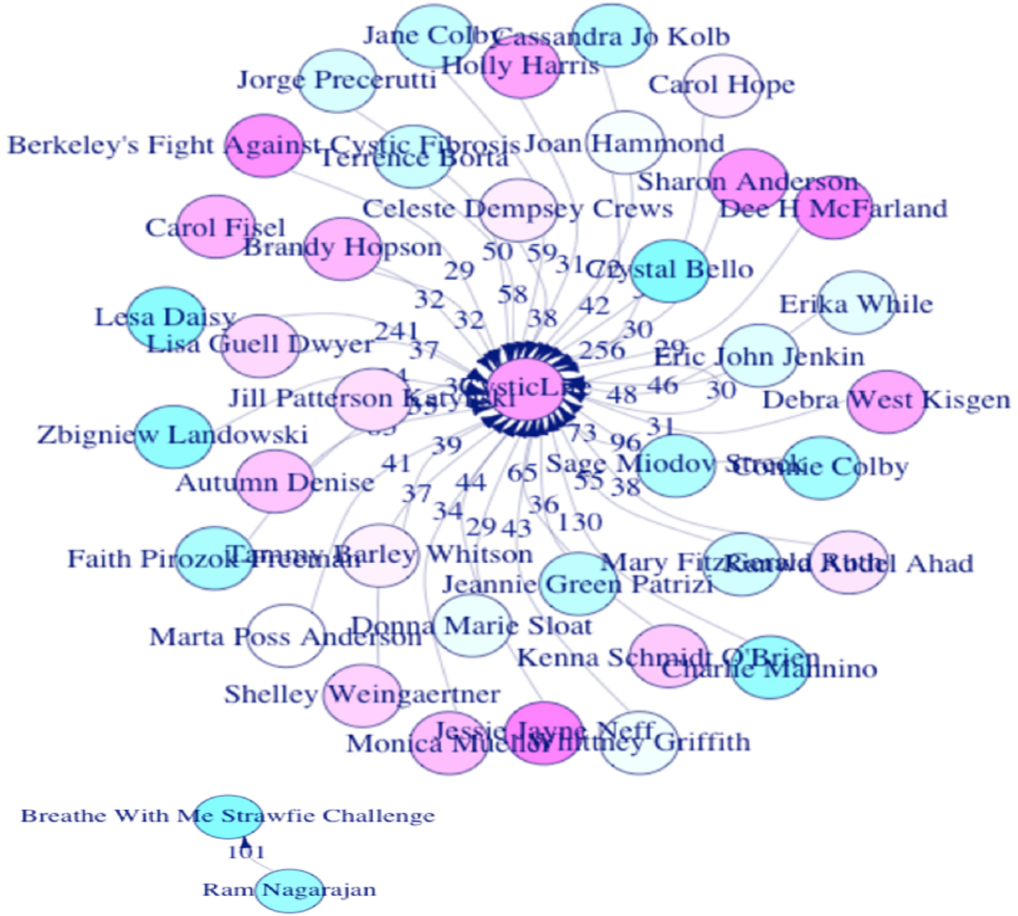
```
lbl = str_c(infl_user_simple$from_name,
      " (", infl_user_simple$weight,")")
pie(infl_user_simple$weight,
    labels = lbl,
    col = varCol,
    main = str_c("Pie Chart for User
       Weights for pageId - ", pageId))
```

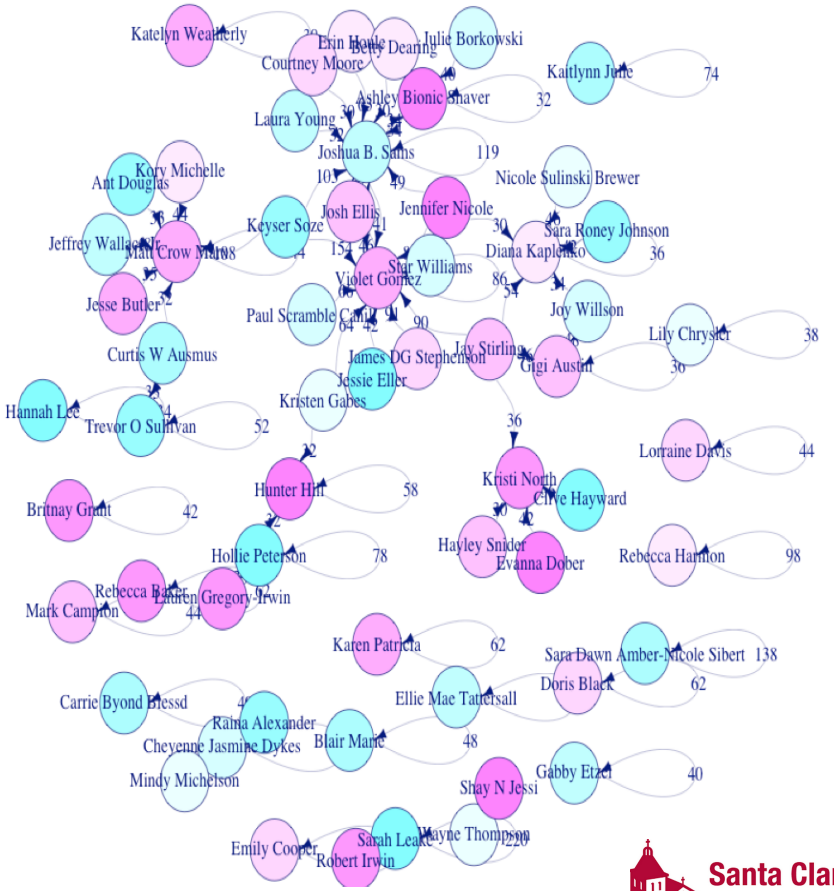




Santa Clara University

Weight Threshold Processed - Network Graphs





Santa Clara University

What are influencers posting on pages?

Cystic Fibrosis Trust

CysticLife.Org

Santa Clara University

Code Snippet for network graph & wordcloud

```r
# Load (DIRECTED) graph from data frame
g_infl <- graph.data.frame(infl_user_simple, directed=TRUE)
# Plot relationshp graph
varCol = cm.colors(length(infl_user_simple$from_name))
# col = rainbow(12)
#plot(g_infl, edge.width=E(g_infl)$weight)
plot.igraph(g_infl,
            edge.arrow.size=0.3, edge.color="navy",
            edge.width = 0.4,
            edge.curved = T,
            edge.label=E(g_infl)$weight,
            vertex.label = V(g_infl)$from_name,
            vertex.label.color = "navy",
            #vertex.size = E(g_infl)$weight/2, vertex.size2 = 40,
            #vertex.size = 60, vertex.size2 = 30,
            vertex.color = varCol,
            vertex.frame.color = "navy",
            #vertex.shape = "rectangle",
            layout=layout.fruchterman.reingold)
```

```r
getTopWordsWordCloud <- function()
{
  print("Preparing to create the wordCloud...")
  raw_corpus = paste(unlist(.GlobalEnv$textCorpus), collapse = '\n')
  raw_corpus = str_replace_all(str_replace_all(raw_corpus, "ystic", ""), "ibrosis", "")
  clean_corpus = str_replace_all(raw_corpus, "[^[:graph:]]", " ")

  ctext = Corpus(VectorSource(clean_corpus))
  ctext = tm_map(ctext,removePunctuation)

  #REMOVE STOPWORDS, NUMBERS, STEMMING
  ctext1 = tm_map(ctext,removeWords,stopwords("english"))
  ctext1 = tm_map(ctext1, removeNumbers)
  tdm <- TermDocumentMatrix(ctext1,control=list(minWordLength=1))
  tdm2 = as.matrix(tdm)
  wordcount = sort(rowSums(tdm2),decreasing=TRUE)
  tdm_names = names(wordcount)

  pal2 <- brewer.pal(8,"Dark2")
  wordcloud(words=tdm_names,freq=wordcount,scale=c(5, .1),min.freq=6,max.words=30, random.order=FALSE,
            random.color=TRUE, rot.per=.14, colors=pal2)
}
```

Few R Coding Learnings to Share

- ➔ Structured Code using custom fns. and configurable parameters

```
assign("filePath", filePath, envir=.GlobalEnv)
path = .GlobalEnv$filePath
```

- ➔ Using Global Variables

```
# MAIN Begin #
tryCatch({...}, warning = function(war)
  {...}, error = function(err)
  {...}, finally = { print("ending main") }) # END tryCatch
# MAIN End #
```

- ➔ Warning and Error Handling by Try Catch. Warning and Error logging - by creating and appending the logs on shared path

- ➔ Creating Customized File name

```
filename = str_c(str_replace_all(str_replace_all(format(Sys.time(),
"%a_%d_%b_%r_%Y"), " ", "_"), ":", "_"),"_Ranking_",pageId)
```

- ➔ Making Colored Word Clouds by using RcolorBrewer::brewer.pal()

```
pal2 <- brewer.pal(8,"Dark2")
wordcloud(words=tdm_names,freq=wordcount,scale=c(5, .1),min.freq=6,
max.words=30, random.order=FALSE, random.color=TRUE, rot.per=.14, colors=pal2)
```

Santa Clara University

Creation of R Package is in progress..

- Facebook Social Graph Influencer Analysis
- FB Marketing Campaign Analysis using Word Clouds

Santa Clara University

R packages used

- library(Rfacebook)
- library(ROAuth)
- library(utility)
- library(rjson)
- library(httr)
- library(httpuv)
- library(SnowballC)
- library(Rook)
- library(stringr)
- library(igraph)
- library(plyr)
- library(gridBase)
- library(NLP)
- library(tm)
- library(wordcloud)
- library(RColorBrewer)

Santa Clara University

contacts

Kanchan Chauhan <kchauhan@scu.edu>
Ram Nagarajan <mramsundar@scu.edu>

Deliverables

- **R source code** file for the complete process of data extraction, data processing, and generating the reporting files.
- **Reporting files:**
  - Public pages + their multi level Public posts CSV and RDA files
  - Influencers report with weights CSV file and a (celebrity flag)
  - Ranking report CSV file
  - **Graphs**:
    - Adjacency matrix network graph with nodes ID
    - Influencing profiles network chart with the computed weight as edges label
    - Weight pie chart
    - Ranking bar chart
    - Word cloud images

Santa Clara University