From Facebook Media Network To Social Network Impacts on Business Metrics

Bentley University

Class: MA710 – Data Mining, Fall 2022

Instructor: Professor Bhaduri Moinak

Group 5: Xuefei Qiao

Yaodan Cui

Yaqiang Liu

Khoa Nguyen

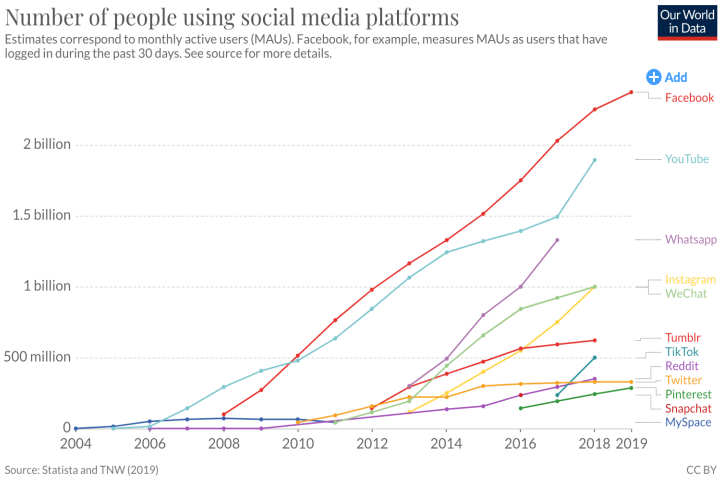
I. INTRODUCTION & MOTIVATION

On July 21, 2022, Simo Kemp released a report, "Digital 2022: July Global Statshot Report", on website www.datareportal.com with a highly interesting figure:

- Total global' s population is about 7.98 billion people
- Global mobile users grew to 5.34 billion or 67%
- Global internet users reached to 5.03 billion or 63.1%
- Social media users grew to 4.70 billion or 59% of the world's total population

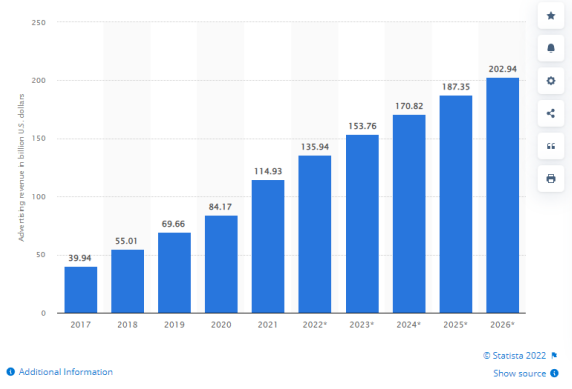
Obviously, the term "social media" or "social network" would make people start thinking about Facebook, YouTube or Instagram, etc.; however, is it much more than just entertainment in reality? A general definition of social media is an interactive and internet-based website and/or applications to connect and/or share contents to other users who are on the same platform application. Research in the past years has been focusing on how consumer behavior has changed over-time in the impacts of social media especially after dot.com and current pandemic COVID-19: one is opening a gate to let people approach to internet and one is forcing people separately in physical contact. People have not formed their friendships only through fiscal contact anymore but also through other social media platforms; business' scale does not hold only at local level but also at global' s: a business has more opportunities to reach its potential customers worldwide by some "clicks" with a low cost. Therefore, which is a fair statement that social media is good or not good for business? Within this paper, we would have a fair answer.

To this point, our group assumes that we agree that social media or social networks has become a crucial part not only in daily life in connecting people but also in economics as well. One of our data sources indicates that "The global social media market grew from $159.68 billion in 2021 to $221.29 billion in 2022 at a compound annual growth rate (CAGR) of 38.6%. ... The social media market is expected to grow to $777.64 billion in 2026 at a



CAGR of 36.9%". The annual growth of about 37% is an impressive number that makes social media become a crucial marketing media tool to many businesses. To narrow the analysis scope, our group introduces Facebook platform as an example of data because of its well-known platform, large number of users, its business model, etc.

Advertising revenues generated by Facebook worldwide from 2017 to 2026(in billion USD)



Facebook's revenue majority comes from advertising in which the company charges its customers who are businesses on how to advertise their contents to their users or their customers. The left-side plot suggests Facebook's revenue and its forecast to fiscal year 2026. Another reason for us to pick Facebook is to learn how a business expands its customer base on the global scale in what costs, which is highly limited in a traditional business model in which business relies heavily on local customers.

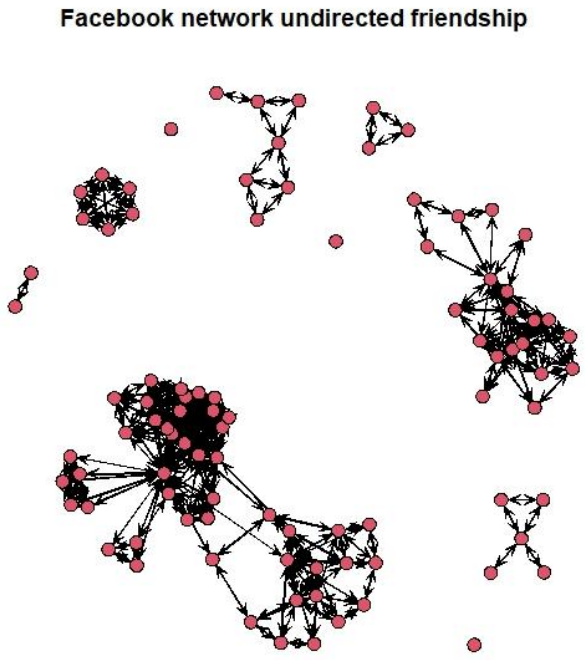
Within this paper we're seeking for the answers to some business questions, such as:

- What is the probability of the "word of mount" given the assumption that A share fully all information to his/her friend B?
- How likely is a given network act to a new person coming to the town? Given the same action to introduce a new product to the market, how likely is that new product to performs?
- How likely is a successful chance to find a person with n^{th} mutual friends through a given sample network? What does it mean if a business can have more information or is able to expand its customers' network?
- If a business finds its target customers, how much should it spend to reach that target?
- How fast a business can target its 100 customers and how much is in budget?

II. EXPLORATORY DATA ANALYSIS

Data used within this paper is from Douglas Lukes's UserNetR package containing nodes' detailed properties and a list of mutual friendships. The property of the network includes "group", "name", "gender", "relationship_status", "friend_count", "mutual_friend_count". There are 93 individual single node names with 323 edge-friendships forming as a simple graph data frame under simple and no loops method that mean there is no repeated node within given data and giving network density measure at 0.0755. As the screenshot suggests the network is formed by 10 separate components with the maximum steps of an efficient way to reach out a person from any given person within a specific component is 3 steps. That means the network has a very high likelihood of people knowing each other if they are from the same component. That also means the community detection is clear to point-out its subcommunity. Since the "distance" between nodes is small, which is 3 steps as discussed above, and large number of components, we expect to see a high value of clustering coefficient, which is 0.666279.

```
> network.size(fb.network)  # network size
[1] 93
> gden(fb.network) # network density
[1] 0.07550257
> components(fb.network) # network components
[1] 10
> max(geodist(component.largest(fb.network,result = "graph"))$gdist)
[1] 3
> # the most efficient steps to communicate from a given node to another within a specific component
> gtrans(fb.network,mode="graph")
[1] 0.6662791
> # clustering coefficient: probability a person to from friendship with his friend within network
>
```

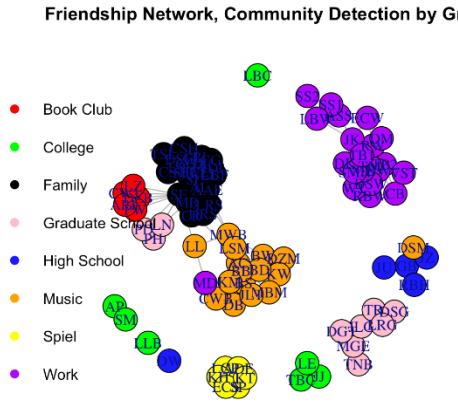


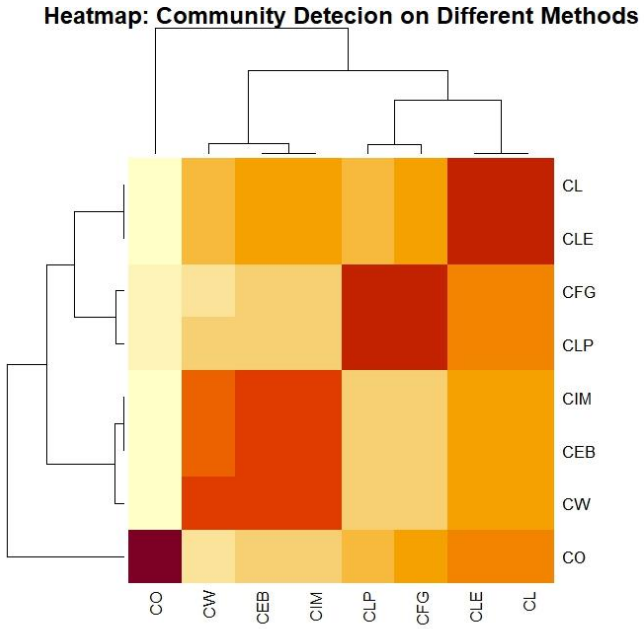
The assortative information presented in table below measures the tendencies to form friendship between 2 persons in which positive sign indicates they are on the same party while negative sign reads different parties within that factor. Assortative coefficient is measured in |1| as:

Facebook Frienship Network Polarization	
Node Properties	**r_a**
name	-0.0184
group	0.9126
gender	0.0127
relationship_status	0.0578
friend_count	-0.0166
mutual_friend_count	0.0969

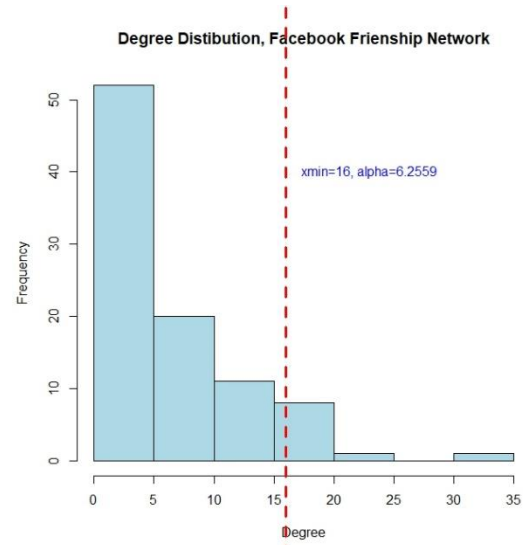
$$-1 \le r_a = \frac{\sum_i fii - \Sigma_i (fij * fji)}{1 - \Sigma_i (fij * fji)} \le 1$$

It's not surprise that the tendencies forming friendship within "group" is extreme high – 0.9126- that we can infer to network influence if a new person comes to the party; however, the feature "gender" is a big surprise when its assortative coefficient presents very low -0.127 - or in another way suggesting insignificant in this network context. Since the assortative coefficient of the "group" is high, we test our community detection classified by "group" to see if it's a good fit which is presented on the right-side. Another test on community detection with different methods to seek for the most agreement among those methods that is presented on the heatmap below.

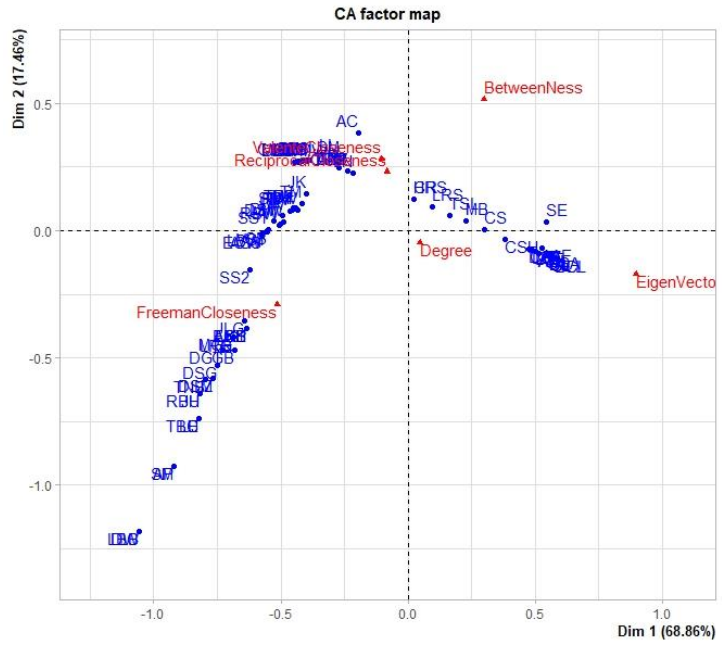




Let's turn our analysis to node details which focus on centrality. The below degree distribution plot clearly suggests with x_min=16 that the node degree distribution is extreme inequality or in another word, node degree distribution for nodes who have more than16 friendships is highly inequality. Alpha value – 6.2559 from fit_power_law() function also suggests extreme inequality, and the node who has the highest degree distribution or most friendship formation is "SE" whose degree distribution is 32 or who has 32 friends within our given network data. Our below code also confirms the most "wealthy" in the network.




```
> E(fb.graph) [303] #SE--AC
+ 1/323 edge from bd387d9 (vertex names):
[1] SE--AC
> get.vertex.attribute(fb.network, "vertex.names")[1] #SE has 32 immediate friends
[1] "SE"
> max(degree(fb.network, gmode='graph'))
[1] 32
>
```



Our most "wealthy" person – SE – also is the center of our network universe in which he is a person in the "middle" and most close to anyone else in our network. Multiple Centrality Assessment method on the left-side plot confirms in both Dim1-axis and Dim 2-axis.

III. THEORY AND METHODS

To answer the business context questions mentioned in prior part, our group applies and uses mainly 2 methods:

- Topological method – a method focusing on to exploits only the network structure, not the node details, such as network density, alpha from power_law with alpha, , node detail degree distribution, etc. For example, for every node pair (i, j), a score function S(i, j) is produced. If it's high, there's a big chance for an edge to be formed friendship between node_i and node_j
- Statistical method – a method aiming to exploits mainly the node details and observations that are common in social networks such as node's gender, node's groups, node's friend_counts, etc., but not so much on the network structure.

There are many different choices have been used for either two methods, but we're focusing on few selections as below:

- Common neighbors
- Preferential attachment
- Jaccard's coefficient
- ERGM—Exponential random graph models
- SIR—Susceptible Infected Recovered

Each method has its owned logic, algorithm, also pros and cons. Topological method, especially Common neighbors, aims to analyze the network structure to understand how many friends a person A has, and the same for a person B to learn how many friends she/he has, then to analyze their common friends to either or not suggest their friendship formation basing on the number of common friends that both person A and person B have. The larger number of common friends they have, the large likelihood chance they would form their friendship. It sounds likely clustering coefficient which is 0.6628 for this given data network – a high number. Regardless of ease in using common neighbor method, the cons of common neighbors is it does not identify the properties of common mutual neighbors that both person A and person B have. Person A and person B may have some common friends, but those common friends maybe formed under unique person A's properties or unique person B's properties that person A does not have or person B has, and vice versa.

To visualize and calculate, preferential attachment and Jaccard's coefficient use below formulars:

For Preferential attachments: $S(i, j) = |N_{(i)}| * |N_{(j)}|$

$$For\ Jaccard's\ coefficient: S_{(i,j)} = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

From the formular and topological method's description, these methods do not investigate the node's properties, such as gender, group, relationship_status, etc. but rather to the network structure. In addition, these methods assume the network structure itself does not change over-time that means we are unable to test what will happen if a new actor comes to the "town". On the other hand, topological method is known as ease to use and is able to solve some social network problems in which the nodes' details are unknown or credential to be released.

To solve topological's cons, the statistical method, which are ERGM and SIR/SIS model, can deal with changes of network structure, nodes' detailed properties, etc. Like topological, ERGM calculates the probability of a link-edge formation between $node_i$ and $node_j$ as:

$$P(y_{ij}|Y_{ij}^C) \sim exp\{\sum_{k=1}^{K} \theta_k z_k(y)\}$$

$$log\{\frac{p}{1-p}\} = \beta_0 + \beta_1 x \iff p = \frac{1}{1 + exp(-\beta_0 - \beta_1 x)}$$

ERGM also release its coefficient so that we can interpret the roll of each property in the model that means we can either select or not select that property to the model based on its significant p-value. In addition, ERGM allow us to justify or emphasize how node's properties are presented in the model. For example, for the same quantitative property, we can either use argument nodecov() as general different values or justify argument to absdiff() to magnitude the difference feature between nodes. Another example for qualitative property is used with either argument nodefactor() to classify nodes' property matching with given condition, or argument nodematch() to emphasize whether the two nodes agree on their attribute. Moreover, other criteria such as AIC/BIC, precision or recall ratio, etc. are presented as a part of function summary(). All these figures will help tremendously to final decision - which model perform better on given data set. Finally, an expanding SIR model will also help to answer other questions relating to marketing budget with trial-error of beta and gamma.

IV. MODELING

To select a model over other models, we use the list of below of 7 criteria with equal weighting on any criterion to simplify our selection; however, depend on context and business questions, weighting may apply to these criteria. With our given data set and context, the more check points are the better score for model selection.

1) Significant variables' co-efficient

2) Low AIC or BIC to prevent overfitting

3) Estimate the density / diameter accurately

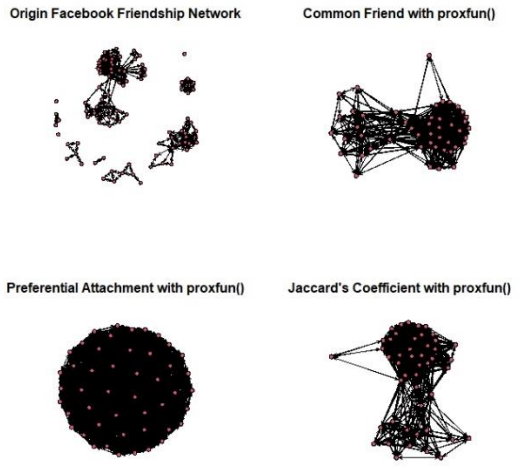
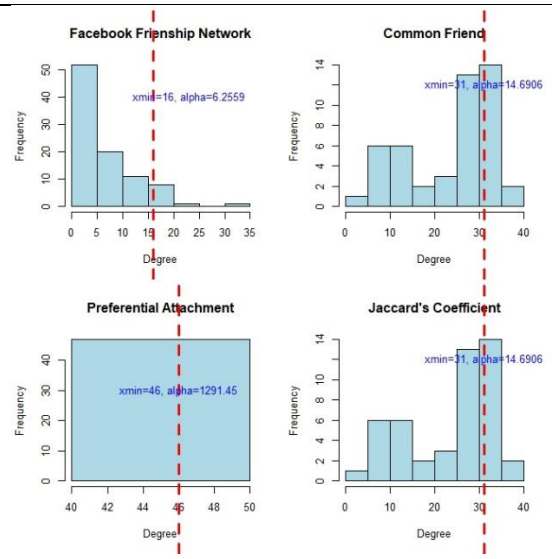
4) Estimate the "wealthy inequality" correctly

5) Estimate community detection correctly

6) Good balance ratio dyad and triad

7) High precision and recall ratios

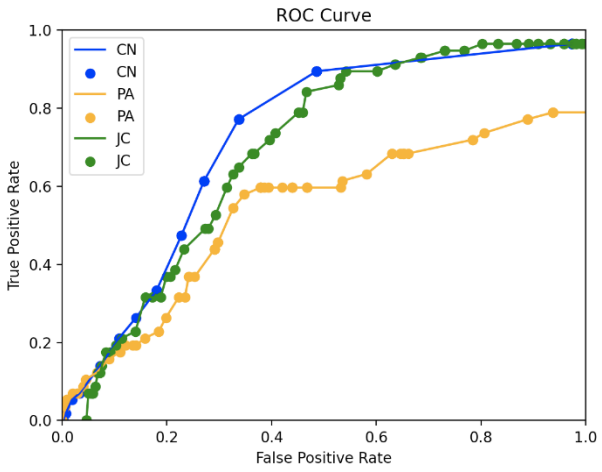
For topological method including Common Neighbors, Preferential Attachment, and Jaccard's Coefficient, we're limited by the model nature which does not provide variable's coefficient, AIC/BIC measurement, community detection; therefore, we only use the other criteria to select a better model. The plots and table below summarize some network's link formation.

Figure: Origin Friendship Network versus Topological with proxfun() for Friendship Prediction	*Figure: Origin Friendship Network versus Topological with proxfun() for Node Degree Distribution*
	

Topological - Model Selection				
	Origin Network	Common Friends	Preferential Attachment	Jaccard's Coefficient
Density	0.0755	0.5198	1	0.5198
Diameter	3	3	1	3
Clustering Coefficient	0.6663	0.8454	1	0.8454
x_min	16	31	46	31
alpha	6.2558	14.6906	294.4511	14.6906

From 2 plots and summarize table to compare 3 methods: common neighbors, preferential attachment, and jaccard's coefficient with the origin Facebook friendship network, it's not hard to eliminate method preferential attachment; however, common neighbors and jaccard's coefficient result the same outputs, so that we'll take another step to justify our argument with the last criterion – precision ratio and recall ratio – that focus mainly to False Positive, False Negative, True Positive and True Negative.

Given the original network containing 10 components, we pick a component with the strongest connection which has 227 edges. We first split the strongest component to training (170 edges and 47 nodes) and test set (57 edges and 47 nodes) according to the random 75% and 25% edges rule, then we get N (Negative) =47*46/2=57 and P (Positive)=57 from the test set. Then, we apply threshold value (start=0, end =1, step=0.01) on training set to determine friendship formation between nodes, and to compare with actual friendship formation to calculate FP – False Positive and TP – True Positive. Lastly, we calculate TN – True Negative and FN – False Negative based on the formular TN+FP=N and FN+TP=P to get the TPR – True Positive Rate and FPR – False Positive Rate. The plot AUC-ROC below shows how these methods perform. Among 3 methods, Common Neighbors – CN method perform better than the other 2 methods which are Preferential attachment and Jaccard's coefficient as it covers more arear under curve.



With the last criterion add-in to compare and select winner model, common neighbor method wins over jaccard's coefficient in the threshold range from 0.20 to 0.60, then even in range

0.60 to 0.85, finally to even in the last threshold 0.85 – 1.00. Taking AUC, we can conclude that common neighbors is a winner. Moreover, even in a case, AUC does not significant differ from common neighbors to jaccard's coefficient, we prefer common neighbor method due to its simplicity.

Due to business context and target business questions, topological method is unable to answer some business questions, we move on to statistic method to aim answering other business questions. With ERGM modeling method, we apply the 7 criteria checklists above to select a better model. Below are 6 screenshots capturing our 6 models' summaries with trial-error starting by ESGM null model to add more node attributes. The summary() helps to identify the first 2 criteria: significant coefficient, and AIC/BIC number.

```
> summary(model0)
Call:
ergm(formula = fb.network ~ edges)

AIC: 2292  BIC: 2298
```

```
> summary(model1)
Call:
ergm(formula = fb.network ~ edges + nodematch("fb.df.group",
    diff = TRUE))

AIC: 1080  BIC: 1137
```

```
> summary(model2)
Call:
ergm(formula = fb.network ~ edges + nodematch("fb.df.group",
    diff = TRUE) + nodefactor("fb.df.relationship_status"))

AIC: 1076  BIC: 1153
```

```
> summary(model3)
Call:
ergm(formula = fb.network ~ edges + nodematch("fb.df.group",
    diff = TRUE) + nodefactor("fb.df.relationship_status") +
    nodecov("fb.df.mutual_friend_count"))

AIC: 838.2  BIC: 920.8
```

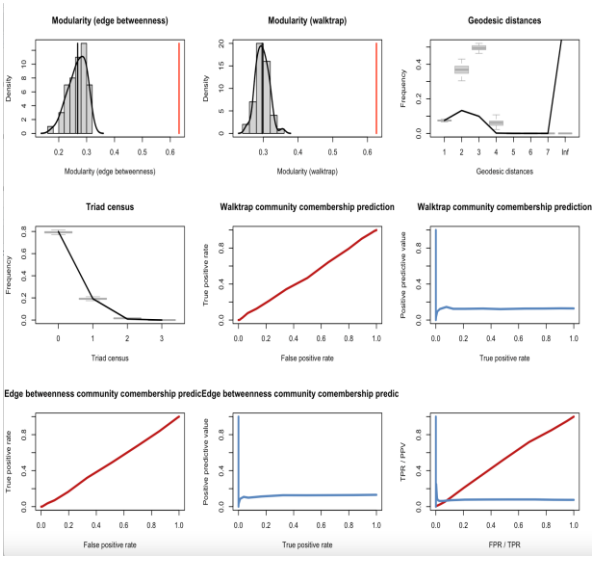
```
> summary(model4)
Call:
ergm(formula = fb.network ~ edges + nodematch("fb.df.group",
    diff = TRUE) + nodefactor("fb.df.relationship_status") +
    nodecov("fb.df.mutual_friend_count") + absdiff("fb.df.friend_count"))

AIC: 835.4  BIC: 924.5
```

```
> summary(model5)
Call:
ergm(formula = fb.network ~ edges + nodematch("fb.df.group",
    diff = TRUE) + nodecov("fb.df.mutual_friend_count") + absdiff("fb.df.friend_count"))

AIC: 829.6  BIC: 899.6
```

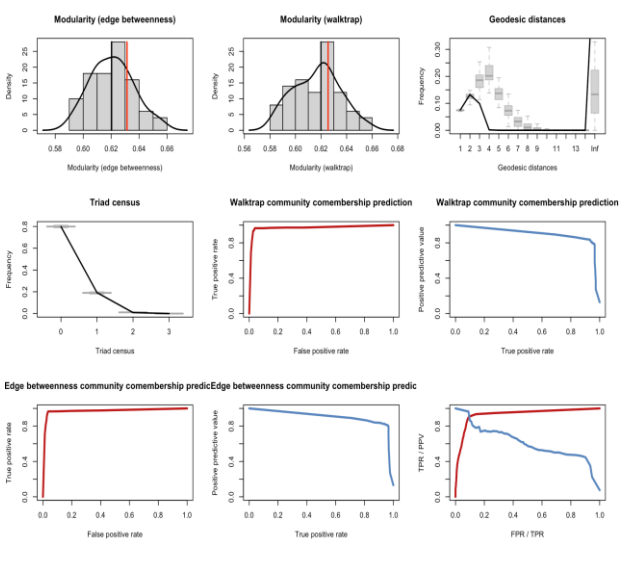
```
> summary(model5)
Call:
ergm(formula = fb.network ~ edges + nodematch("fb.df.group",
    diff = TRUE) + nodecov("fb.df.mutual_friend_count") + absdiff("fb.df.friend_count"))

Maximum Likelihood Results:

                                  Estimate Std. Error MCMC % z value Pr(>|z|)
edges                           -1.017e+01  6.459e-01      0 -15.749   <1e-04 ***
nodematch.fb.df.group.Book Club  2.430e+01  7.587e+02      0   0.032   0.9745
nodematch.fb.df.group.College    8.336e+00  8.223e-01      0  10.137   <1e-04 ***
nodematch.fb.df.group.Family     4.843e+00  3.251e-01      0  14.898   <1e-04 ***
nodematch.fb.df.group.Graduate School  8.062e+00  6.514e-01      0  12.376   <1e-04 ***
nodematch.fb.df.group.High School  8.831e+00  9.094e-01      0   9.710   <1e-04 ***
nodematch.fb.df.group.Music      7.326e+00  5.173e-01      0  14.162   <1e-04 ***
nodematch.fb.df.group.Spiel      2.478e+01  6.416e+02      0   0.039   0.9692
nodematch.fb.df.group.Work       6.717e+00  4.667e-01      0  14.392   <1e-04 ***
nodecov.fb.df.mutual_friend_count  2.087e-01  1.640e-02      0  12.728   <1e-04 ***
absdiff.fb.df.friend_count      -8.810e-04  4.037e-04      0  -2.182   0.0291 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 5930.6  on 4278  degrees of freedom
 Residual Deviance:  807.6  on 4267  degrees of freedom

AIC: 829.6  BIC: 899.6  (Smaller is better. MC Std. Err. = 0)
```

As we can see above, model5 has the lowest AIC/BIC, 529.6 & 899.6 respectively, and all significant coefficients. The difference between "model5" and "model4" is to remove an insignificant attribute – nodefactor("relationship_status"). Among other "model0", "model1", "model2" and "model3", "model4" has the lowest AIC/BIC; therefore, "model5" is selected over all other models. A goodness-of-fit diagrams is plotted below to reference.

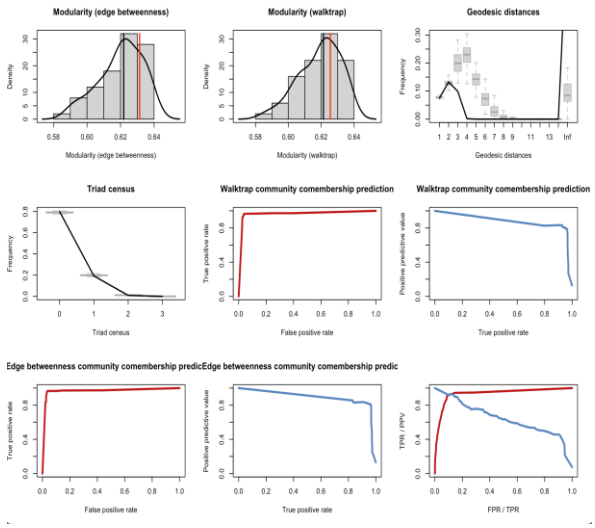
Model0 – Goodness of Fit



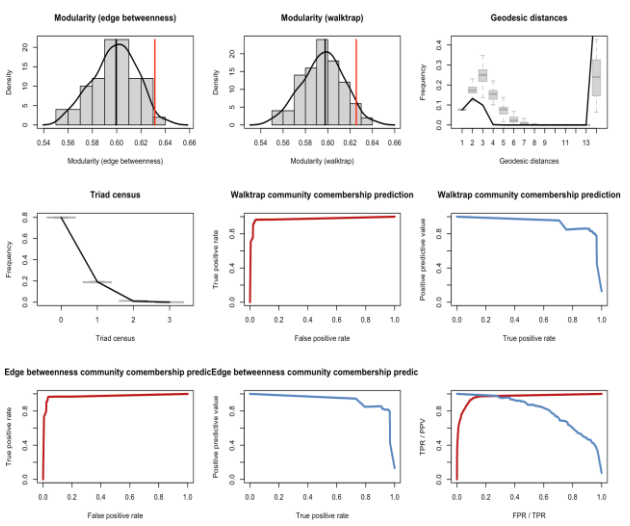
Model1 – Goodness of Fit



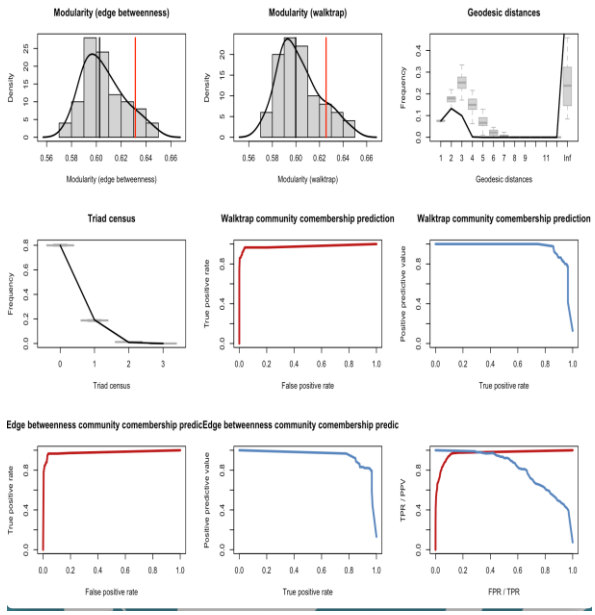
Model2 – Goodness of Fit



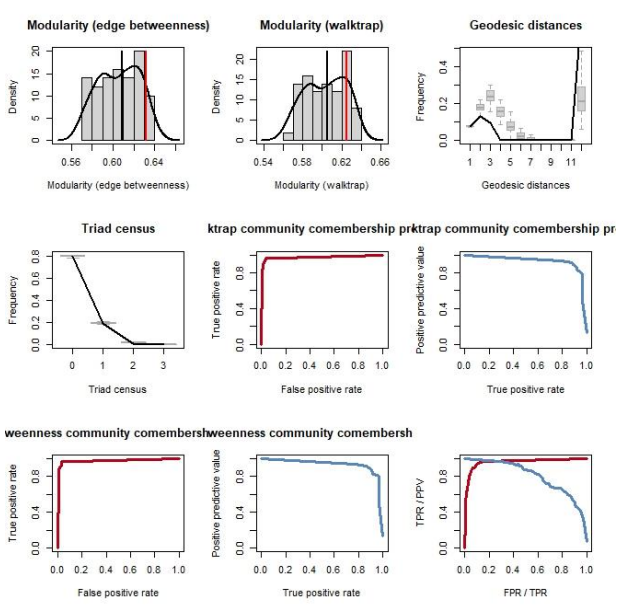
Model3 – Goodness of Fit



Model4 – Goodness of Fit



Model5 – Goodness of Fit



Adding to significant coefficient and low AIC/BIC, "model5" performs accurately in the balance of triad and dyad ratios which is plot 4, and high accuracy of precision and recall ratio which is plot 5 to plot 9. However, it does not perform well community detection, density, diameter,

and "wealthy inequality" which is plot 1 to plot 3. Overall, ERGM – model 5 is best ERGM model we select.

Model Name	*Model0*	*Model1*	*Model2*	*Model3*	*Model4*	*Model5*
AIC	2292	1080	1076	838.2	835.4	829.6
BIC	2298	1137	1153	920.8	924.5	899.6
# of Significant Predictors out of All Predictors	1 out of 1	2 out of 2	3 out of 3	3 out of 4	4 out of 5	4 out of 4
Modularity Absolute Difference (edge betweenness)	0.36	0.01	0.01	0.035	0.035	0.025
Modularity Absolute Difference (walktrap)	0.34	0.0025	0.005	0.024	0.022	0.022
Geodesic Distance	Bad	Shows the trend	Shows the trend	Shows the trend	Show the trend	Show the trend
ROC Plot	Bad	Very fast to reach top left	Very fast to reach top left	Very fast to reach top left	Very fast to reach top left	Very fast to reach top left

V. HOW TO APPLY – BUSINESS QUESTIONS

1) How hard to find a random person who has 30 friends or only 2 friends within given network?

Regardless of a particular network with given a large number of nodes or not, if that network has power-law with alpha of α, the probability of finding an actor or a person with particular k number of friends would be defined as: $Prob(X = k) \sim \frac{1}{k^{\alpha}}$

Therefore, if a node has immediate 2 friends and its entire network was given alpha- α=6.6255, the probability to find that person is:

$$Prob(X = k) \sim \frac{1}{k^{\alpha}} = \frac{1}{2^{6.6255}} = \frac{1}{98.53} = 1.309\%$$

The same idea applies to find a person who has 30 immediate friend within that network is:

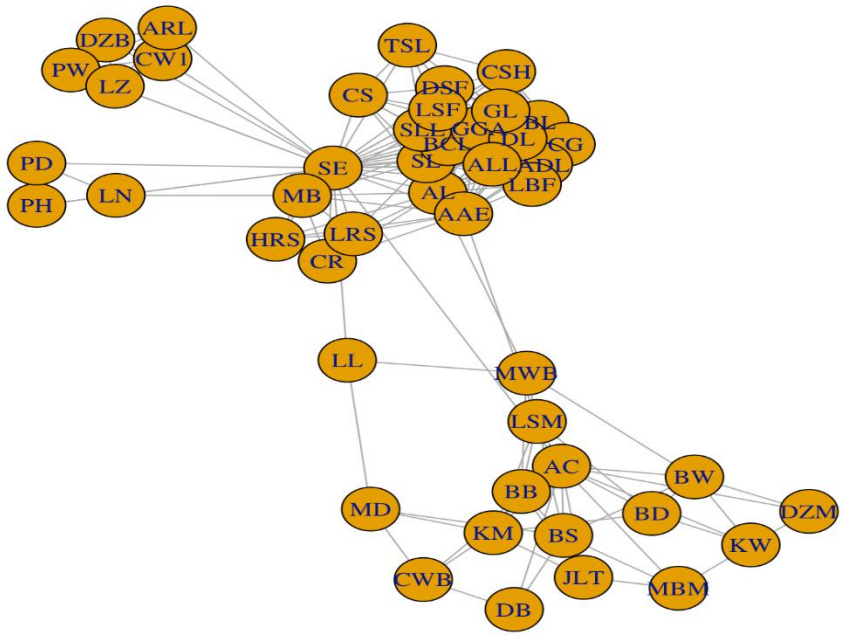
$$Prob(X = k) \sim \frac{1}{k^{\alpha}} = \frac{1}{30^{6.6255}} = \frac{1}{6056692283} = 5.7624e^{-10}$$

Below is also a R-code to confirm our hand calculating above.

```
> print("Probability to find someone who has 2 friend within given network, alpha = 6.6255:")
[1] "Probability to find someone who has 2 friend within given network, alpha = 6.6255:"
> 1 /(2^6.255)
[1] 0.01309355
> print("Probability to find someone who has 30 friend within given network, alpha = 6.6255:")
[1] "Probability to find someone who has 30 friend within given network, alpha = 6.6255:"
> 1 /(30^6.255)
[1] 5.762442e-10
>
```

2) Within a given strongest component as defined above, if person KW want to reach to a person who has the most friend, how much in budget does person KW need?

Given that the diameter of the original network is 3 as described below, the maximum number of steps that person KW want to reach to any one within that component is also 3. If person KW needs more than 3 steps to reach out to any other person, that approach is considered not efficient. From the code and plot below, we can identify that person SE is a person who has the most friend within the network. This information also confirmed in prior part that we identify centrality.



Applying the adjacency matrix extracting from given data, if 2 persons, SE and KW, don't form their friendship, the adjacency matrix corresponding to these 2 person will be 0 or A[SE ,KW] = 0. Otherwise, A[SE,KW] > 0 indicates that these 2 persons SE and KW form their friendship. The last code on the right-side calling for matrix2 = matrix1 * matrix1, suggests A[SE, KW] = 1 which means within budgets of 2 steps, person KW is able to reach person SE who has the most friends in given network.

```
##Q2
strong_graph #47 nodes, and 227 edges
strong_network=asNetwork(strong_graph)
plot(strong_graph, labels=TRUE)
degree(strong_network, gmode='graph')
which(degree(strong_network, gmode='graph')==max(degree(strong_network, gmode='graph')))
max(degree(strong_network, gmode='graph'))
get.vertex.attribute(strong_network, "vertex.names")[1] #SE has 32 immediate friends


get.vertex.attribute(strong_network, "vertex.names") #SE: index=1, KW: index 13
matrix1=get.adjacency(strong_graph)
matrix1
matrix2=matrix1 %*% matrix1
matrix2[1,13] #1
```

3) Without being given any node detail information, how hard a person would from friendship with another person?

```
538  ##Q3
539  model0 = ergm(fb.network~edges)
540  summary(model0)
541  manual.prob=1/(1+exp(2.50508))
542  manual.prob #7.55%
543
```

545:1 # (Untitled)

Console | Terminal | Background Jobs

R 4.2.1 · ~/Desktop/MA710_Data_Mining/Group_Project/

```
> summary(model0)
Call:
ergm(formula = fb.network ~ edges)

Maximum Likelihood Results:

      Estimate Std. Error MCMC % z value Pr(>|z|)
edges -2.50508    0.05787      0  -43.29   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 5931  on 4278  degrees of freedom
 Residual Deviance: 2290  on 4277  degrees of freedom

AIC: 2292  BIC: 2298  (Smaller is better. MC Std. Err. = 0)
> manual.prob=1/(1+exp(2.50508))
> manual.prob
[1] 0.07550282
```

The code above indicates how model0 or called null model is built by applying ergm() function to build a statistic model. In the model's summary, the coefficient values at -2.50508 with its significant p-value at < $1e^{-4}$. The significant p-value suggests that coefficient value is stable with given data and is good to calculate the probability of forming friendship between 2 peoples.

$$log\{\frac{p}{1-p}\} = \beta_0 + \beta_1 x \iff p = \frac{1}{1 + exp(-\beta_0 - \beta_1 x)}$$

$$Prob \sim \frac{1}{1+e^{-(\beta 0+\beta 1)}} = \frac{1}{1+e^{2.50580}} = \frac{1}{13.253} = 7.545\%$$

The probability for 2 persons forming their friendship is about 7.545% with our given network.

4) How likely does 2 people becoming friend given some properties if one user is in College club who has about 26 mutual friends with the author, and a second user is in Music club who has 6 mutual friends with the author, and given that they both have absolute difference on friend count as 200?

```
> summary(model5)
Call:
ergm(formula = fb.network ~ edges + nodematch("fb.df.group",
    diff = TRUE) + nodecov("fb.df.mutual_friend_count") + absdiff("fb.df.friend_count"))

Maximum Likelihood Results:

                                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges                             -1.017e+01  6.459e-01      0 -15.749   <1e-04 ***
nodematch.fb.df.group.Book Club    2.430e+01  7.587e+02      0   0.032   0.9745
nodematch.fb.df.group.College      8.336e+00  8.223e-01      0  10.137   <1e-04 ***
nodematch.fb.df.group.Family       4.843e+00  3.251e-01      0  14.898   <1e-04 ***
nodematch.fb.df.group.Graduate School  8.062e+00  6.514e-01      0  12.376   <1e-04 ***
nodematch.fb.df.group.High School  8.831e+00  9.094e-01      0   9.710   <1e-04 ***
nodematch.fb.df.group.Music        7.326e+00  5.173e-01      0  14.162   <1e-04 ***
nodematch.fb.df.group.Spiel        2.478e+01  6.416e+02      0   0.039   0.9692
nodematch.fb.df.group.Work         6.717e+00  4.667e-01      0  14.392   <1e-04 ***
nodecov.fb.df.mutual_friend_count  2.087e-01  1.640e-02      0  12.728   <1e-04 ***
absdiff.fb.df.friend_count        -8.810e-04  4.037e-04      0  -2.182   0.0291 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 5930.6  on 4278  degrees of freedom
 Residual Deviance:  807.6  on 4267  degrees of freedom

AIC: 829.6  BIC: 899.6  (Smaller is better. MC Std. Err. = 0)
> print("The probability for 2 persons with given detailed node's properties is:")
[1] "The probability for 2 persons with given detailed node's properties is:"
> 1/ (1 + exp(-(-1.017e+01 + 8.336 + 7.326 + 0.2087*6 + 0.2087*26 - 8.810e-04*200)))
[1] 0.9999938
>
```

Apply ERGM model to calculate the probability of these 2 people to form their friendship as:

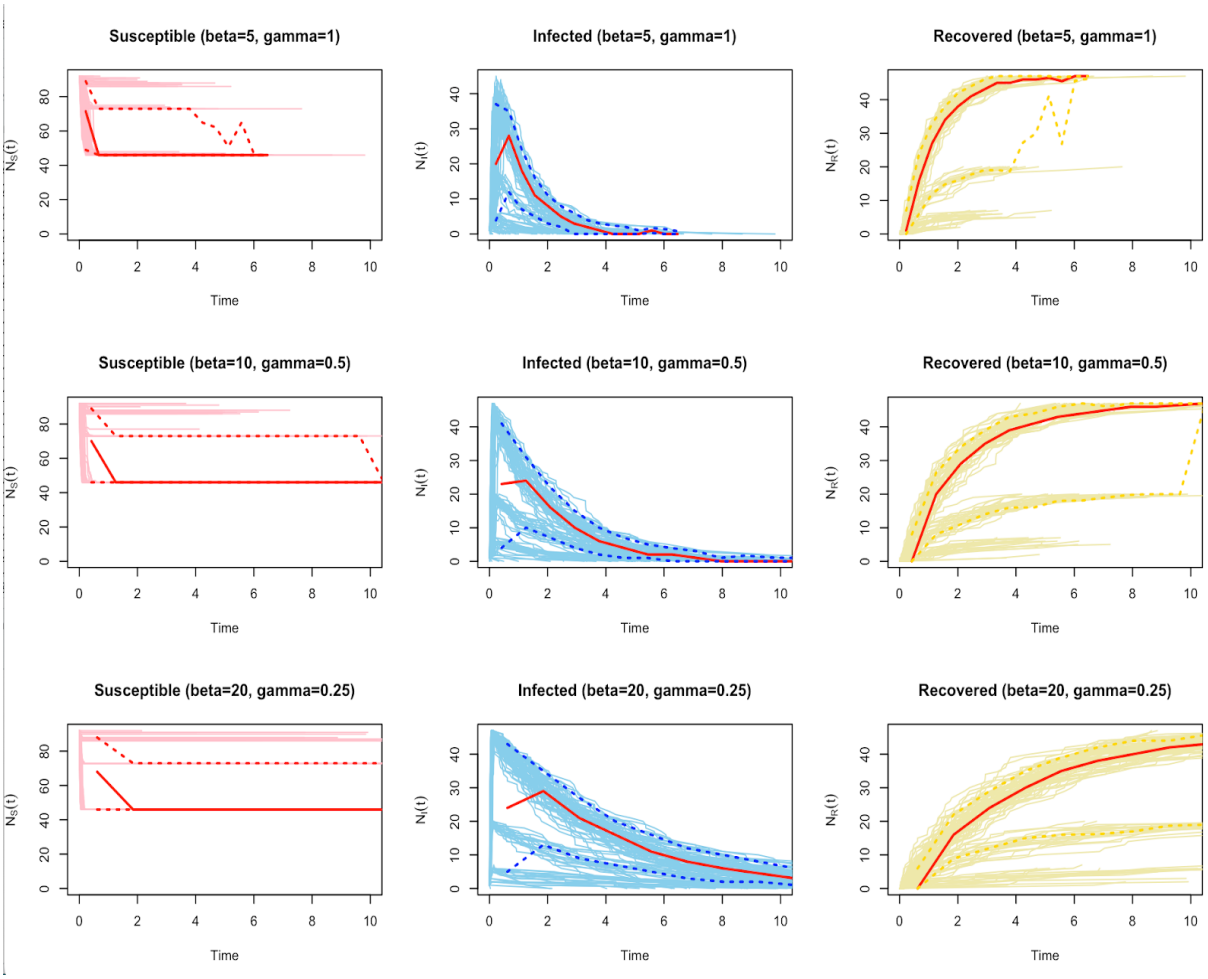
$$P(y_{ij}|Y_{ij}^{C}) \sim exp\{\sum_{k=1}^{K} \theta_k z_k(y)\}$$

$$log\{\frac{p}{1-p}\} = \beta_0 + \beta_1 x \Longleftrightarrow p = \frac{1}{1+exp(-\beta_0 - \beta_1 x)}$$

Based on the R code, these two users have a very high chance (close to 100%) will become friends on Facebook social media network.

5) Given social network information stable with either no actor to join or leave the community, and with a budget of 10 flyers, how many people expected to be inflected at either time 1 or time 2?

Because there is no person to join or leave the social network community, and with a fixed budget of 10 flyers, plot below presents trial-error number of people being infected by time 1 or time 2.



From the plot, at either time 1 or time 2, the maximum number of infected persons is about 30 with either pair of beta=5 and gamma=1 or pair of beta=20 and gamma=0.25. However, if we have larger budget, for example, budget is 20 flyers, we can expect more people infected in time 1 and time 2 with the same beta and gamma values.

VI. CONCLUSIONS & FUTURE WORK

In our analysis, ERGM is a great statistical method to solve our business questions. For all ERGM model we built, model 5 (formula = fb.network ~ edges + nodematch("fb.df.group", diff = TRUE) + nodecov("fb.df.mutual_friend_count") + absdiff("fb>df.friend_count")) has the lowest AIC/BIC, 829.6 & 899.6. It performs great in testing of goodness of fit in which it reaches top left fastes in plot#7 and plot 9, False Positive rate vs. True Positive rate, fairly replicates Geodesic Distance trend, plot 3, and accurately presents Triad Census information

We improve a lot progress in our analysis, but there are still some limitations and questions that need to be solved and improved. The main limitation is dataset. In our dataset, it didn't include multiple layers of information. Multilayer data can help us understand how friendship is formed under different environment such as workplace, or physical community and which is the strongest attributes within these layers. Person "SE" is the most popular in our given network, how she/he performs on other environment, or attribute "group" is still the strongest assortative information or attribute "gender" is the weakest attribute to detect community. The answers for these questions are crucial if we're launching new products to consider with a fixed budget.

Also, our dataset lacks data at different times. In another word, we have the network information for only one moment. Our analysis does not explore how the network changes over time with the impact of either social influence or social selection. For example, with a budget problem, we can expand SIR model to SIS model to apply assumption number of people joins the network or leave the network. Another example, if a business would like to target a certain number of customers within a period of time, how the customer or potential customer changes, and how a person interests in the existing product changing over time. Besides, questions like "social influence" or "network selection" influencing "wealthy" people" or whether "wealthy" level will change over time does not solve in our report. These contain many valuable information for businesses. We might need to consider these improvements in future analyses.

REFERENCES

Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. New York, NY: John Wiley & Sons, 2017.

"Social media," Wikipedia, last modified December 10, 2022, https://en.wikipedia.org/wiki/Social_media

"The Impact of Social Media in the 21st Century." StudyMoose, Nov 18, 2020. Accessed December 12, 2022. http://studymoose.com/the-impact-of-social-media-in-the-21st-century-essay

"Impact of Covid-19 on consumer behavior: Will the old habits return or die?" Jagdish Seth, Accessed June 4, 2020. https://www.sciencedirect.com/science/article/pii/S0148296320303647

Sajal Kohli, Björn Timelin, Victor Fabius, Sofia Moulvad Veranen, "How COVID-19 is changing consumer behavior –now and forever." https://www.mckinsey.com/~/media/mckinsey/industries/retail/our%20insights/how%20covid%2019%20is%20changing%20consumer%20behavior%20now%20and%20forever/how-covid-19-is-changing-consumer-behaviornow-and-forever.pdf

Statista Research Department, 2022, Global Facebook advertising revenue 2017-2026. https://www.statista.com/statistics/544001/facebooks-advertising-revenue-worldwide-usa/#:~:text=In%202021%2C%20Facebook%20generated%20nearly,of%20the%20global%20ad%20revenue