Introduction to Business Analytics Assignment Report

Hoa Giang, Han Mao, Niccolo Valerio, Neha Sharma, Jim Leach

2 November 2015

People picking

Problem

For this assignment the group was tasked wit using network data collected from the Business Analytics course to pick three, five-person teams for (1) design; (2) advocacy; and (3) implementation of graduation week plans.

The catch

The picks were subject to two constraints:

- *Capacity*: Each team must have only five people, and the same individual could not be in more than one team; and
- *Chemistry*: The picks came with a budget. Each team could only use a maximum of 30 "visibility points" (referred to as VPs). These VPs were a proxy for popularity and were derived from a network set up to mimic the social structure of the students.

Picking teams and this document

The team-picking exercise was carried out and the results presented to the rest of the course. This document, therefore, presents the results of some more detailed analysis that was conducted as part of the assignment.

Four questions were assigned that facilitated further exploration and understanding of the networks. The responses to these questions are presented in this document.

Assignment Responses

The assignment was competed using the R language. As part of this, a number of additional packages were used for this assignment:

```
library(MASS)
library(lsa)
library(igraph)
library(readxl)
library(dplyr)
library(magrittr)
library(tidyr)
library(ggplot2)
library(knitr)
library(broom)
```

The data were read in to R and combined in to a list of data frames for ease of computation.

1 - Regressions

Initially, two functions were defined to aid the extraction of the necessary statistics. The first was used to extract the in-degrees centrality value for all nodes in a network, and the second converted the resulting list object to a tidy data frame for ease of processing.

Next, using these functions, the in-degrees centrality was calculated for each of the four networks and the results combined in to a tidy data frame.

An exploratory plot was created to understand the distributions of the in-degrees centrality accross the four networks (see Appendix 1). It was determined that negative binomial regression (a special form of Poisson regression) would probably be most suitable for the problem (as the conditional variances of the distributions were much larger than the conditional means). The models were created and the regression coefficients summarised below.

Table 1: Regression coefficients from negative binomial regression of design picks on social popularity

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.7701438	0.1650958	4.664829	3.1e-06
popularity	0.1098004	0.0270405	4.060583	4.9e-05

Table 2: Regression coefficients from negative binomial regression of implementation picks on social popularity

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.6687405	0.1715936	3.897235	9.73e-05
popularity	0.1484542	0.0277840	5.343163	1.00e-07

Table 3: Regression coefficients from negative binomial regression of advocacy picks on social popularity

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.7365648	0.1804607	4.08158	4.47e-05
popularity	0.1243884	0.0299906	4.14758	3.36e-05

It was seen that in all three models there is a statistically significant (at the 1% level) relationship between the in-degrees centrality in the Albert Hall network (a proxy for general popularity) and the in-degrees centrality in the other three networks.

The *Albert Hall popularity* coefficient in each regression value can be interpretted as:

- For the *design* network: a one-unit increase in in-degree centrality in the Albert Hall network resulted in an increase in the expected log count of the design in-degrees centrality by 0.11 [0.056, 0.165];
- For the *implementation* network: a one-unit increase in in-degree centrality in the Albert Hall network resulted in an increase in the expected log count of the implementaion in-degrees centrality by 0.148 [0.092, 0.207]; and
- For the *advocacy* network: a one-unit increase in in-degree centrality in the Albert Hall network resulted in an increase in the expected log count of the advocacy in-degrees centrality by 0.124 [0.063, 0.188].

2 - Cosine Similarity and flexibility

Cosine similarity can be used to measure the similarity between two (or more) vectors of 0's and 1's. Treating each individuals' picks in a given network in this way allows the similarity of their picks accross the four networks to be determined. Extended from this, a *flexibility score* has been developed.

Given the four "pick-vectors" for each individual accross each network, the cosine similarity between all combinations of these vectors has been determined (i.e. vector 1 with vectors 2, 3, and 4; vector 2 with vectors 1, 3, and 4 etc).

After doing this (and ignoring the comparison of a pick vector with itself), the average (mean) value of the cosine similarity score was calculated to give an approximate, single-value measure for each individual.

It should be noted that given that the `cosine` function measures *similarity*, a lower value actually represents a higher flexibility. Therefore, the cosine similarity was subtracted from one to give an overall flexibility score where a higher value indicates a more flexible individual. As such, the flexiblity is simply one minus the mean cosine similarity for all pick-vector to pick-vector comparisons for each individual.

Table 4: Flexibility score and Z-value for all 57 individuals in the class

ID	Flexibility	Z
24	1.0000000	1.4292731
29	1.0000000	1.4292731
53	1.0000000	1.4292731
39	0.9166667	1.1054986
44	0.9166667	1.1054986
9	0.9087129	1.0745959
46	0.9087129	1.0745959
42	0.8333333	0.7817241
52	0.8333333	0.7817241
50	0.8087129	0.6860665
38	0.7979379	0.6442026
6	0.7958759	0.6361908
43	0.7800868	0.5748457
13	0.7530494	0.4697975
30	0.7500000	0.4579496
34	0.7500000	0.4579496
36	0.7500000	0.4579496
49	0.7500000	0.4579496
32	0.7418011	0.4260945
8	0.7261387	0.3652416
40	0.6924258	0.2342570
10	0.6666667	0.1341751
11	0.6666667	0.1341751
31	0.6666667	0.1341751
37	0.6666667	0.1341751
51	0.6666667	0.1341751
55	0.6666667	0.1341751
25	0.5833333	-0.1895994
48	0.5833333	-0.1895994
1	0.5763932	-0.2165638
12	0.5000000	-0.5133739
3	0.4836022	-0.5770841
45	0.4531260	-0.6954932

ID	Flexibility	Z
16	0.4522774	-0.6987900
14	0.4309644	-0.7815975
2	0.4226497	-0.8139024
17	0.4166667	-0.8371484
23	0.3333333	-1.1609229
28	0.3131133	-1.2394837
35	0.0833333	-2.1322464
56	0.0833333	-2.1322464
54	0.0527864	-2.2509302
15	0.0000000	-2.4560209
4	NA	NA
5	NA	NA
7	NA	NA
18	NA	NA
19	NA	NA
20	NA	NA
21	NA	NA
22	NA	NA
26	NA	NA
27	NA	NA
33	NA	NA
41	NA	NA
47	NA	NA
57	NA	NA

3 - Determining group leaders After assigning five individuals to each team (see supporting presentation material for a description of how this was performed) it was also necessary to pick a team leader. As well as the flexiblity score developed above, additional information was used to assign each member of the team extra flexibility.

Using data from the personality quiz, an individual was assigned:

- one additional point on the flexiblity score for choosing to host a party using two different invitation methods;
- one additional point on the flexiblity score for choosing to use a mix of invitation methods weighted in favour of their preferred method; and
- two additional points on the flexiblity score for choosing to use a 50/50 split mix of invitation methods from their first and second preferences.

After performing this process, the following three IDs were selected to be the lead member of each team.

Table 5: Team leaders based on flexibility scores

Team	ID	Flexiblity Score	Z
Design	52	0.8333333	0.7817241
Implementation	30	0.7500000	0.4579496
Advocacy	42	0.8333333	0.7817241

4 ID Rankings To produce a "cost-benefit" ratio for each network (where the cost is defined as the in-degrees centrality for each node on the Albert Hall network) the picking scores were used. These scores are defined as follows:

- $Design = 0.6Betweenness + 0.3Eigenvectorcentrality + 0.1Closeness$
- $Implementaion = 0.1Betweenness + 0.3Eigenvectorcentrality + 0.6Closeness$
- $Advocacy = 0.4Betweenness + 0.4Eigenvectorcentrality + 0.2Closeness$

Having defined these formulae, the tables below present each ID in each network, along with the associated measure-of-value (which is defined as $\frac{score}{visibility}$).

Table 6: Design network measure of value

ID	Score
42	1.9715924
48	1.9293307
52	1.7755496
55	1.6894286
14	1.3567138
31	1.0766333
35	1.0475553
51	1.0096977
3	0.8123977
8	0.6560915
43	0.5742077
6	0.5343061
23	0.5160858
25	0.4675078
54	0.4381155
30	0.4355138
17	0.4132363
9	0.2947380
15	0.2891459
11	0.2635357
16	0.2616838
37	0.1881685
38	0.1625666
10	0.0731026
44	0.0618393
46	0.0478870
50	-0.0859923
2	-0.0875537
1	-0.1661973
13	-0.2037350
40	-0.2502555
12	-0.3055114
4	-0.3181462
57	-0.3283438
36	-0.3300528
20	-0.3853815
18	-0.3861899
47	-0.4971155
33	-0.5733699

ID	Score
24	-0.6200245
28	-0.6425362
21	-0.6511765
29	-0.6780142
41	-0.7613631
22	-0.7786326
7	-0.8089732
26	-0.8577249
27	-0.8659419
56	-0.9578287
32	-1.1344284
34	-1.1344284
39	-1.1344284
45	-1.1344284
49	-1.1344284
53	-1.1344284
5	NA
19	NA

Table 7: Implementation network measure of value

ID	Score
48	1.5141486
55	1.3229148
42	1.1960673
18	0.9619756
35	0.9383463
51	0.7748445
6	0.7340910
3	0.7233702
25	0.6509160
30	0.5803550
23	0.5563988
57	0.5367165
43	0.5114222
11	0.4011297
39	0.3554677
46	0.3272241
34	0.3077572
52	0.2990011
2	0.2964968
29	0.2674192
50	0.2296320
1	0.2093629
7	0.2022918
22	0.2014916
13	0.1623150
9	0.1412127
21	0.1364949
4	0.1316244
44	0.1295870

ID	Score
40	0.0993263
47	0.0953711
27	0.0833702
38	0.0759528
45	0.0510116
14	0.0395235
31	-0.0205126
15	-0.0326986
10	-0.0456409
20	-0.0639984
41	-0.1076543
37	-0.2008014
8	-0.2151790
49	-0.2311631
33	-0.2341201
36	-0.3258818
24	-0.3362952
54	-0.3965701
28	-0.5382289
16	-0.6262556
12	-1.9437790
53	-1.9665925
17	-1.9682485
26	-1.9970034
32	-1.9970034
56	-1.9970034
5	NA
19	NA

Table 8: Advocacy network measure of value

ID	Score
25	2.1487924
42	1.6906367
50	1.5437607
48	1.2908717
6	0.9554678
35	0.9404589
54	0.9118415
55	0.7906126
18	0.7400452
30	0.7121953
40	0.5834349
52	0.5710575
23	0.5405137
57	0.5009658
10	0.4688547
1	0.3727096
3	0.2719974
51	0.2641880
9	0.2385837

ID	Score
56	0.2285706
46	0.1245581
4	0.1055407
14	-0.0017206
11	-0.1131409
2	-0.1287779
29	-0.1315220
37	-0.2137078
47	-0.2332233
39	-0.2628181
43	-0.2786073
15	-0.2850678
17	-0.2972158
12	-0.3227137
8	-0.3233655
24	-0.3545014
38	-0.3780787
28	-0.3868711
32	-0.4570769
22	-0.4676482
44	-0.4699151
33	-0.5178695
31	-0.5595238
16	-0.5906137
34	-0.6109640
49	-0.6128073
7	-0.6206129
41	-0.6537524
19	-0.7246687
26	-0.7634617
36	-0.7959320
13	-1.4798265
45	-1.4798265
53	-1.4798265
5	NA
20	NA
21	NA
27	NA

Appendices

Appendix One - Exploratory plot of in-degree centrality distributions