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 with using network data collected from the Business Analytics course students 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 students on 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.

The assignment was competed using the R language. The data were read in to R and the `igraph` package (amongst others) was used to perform the network analysis.

Assignment Responses

1 - Regressions

The relationship between the in-degree centrality of each node in the Royal Albert Hall network (the "visibility points") and the in-degree centrality in the remaining three networks was assessed to determine if and how popularity (visibility points) affects the likelihood of being picked elsewhere.

An exploratory plot was created to understand the distributions of the in-degrees centrality for each network (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 (please note that OLS linear regression was also performed and the results summarised in Appendix 2):

Table 1: Regression coefficients from negative binomial regression of design in-degree centrality on visibility points (vp)

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

Table 2: Regression coefficients from negative binomial regression of implementation in-degree centrality on visibility points (vp)

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

Table 3: Regression coefficients from negative binomial regression of advocacy in-degree centrality on visibility points (vp)

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

In all three models there was a statistically significant (at the 1% level) relationship between visibility points and the in-degrees centrality in the other three networks.

The *vp* coefficient can be interpreted as the expected change in the log count of in-degree centrality in the other network(s) given a one-unit change in *vp*.

- For *design* this value was 0.11 [0.056, 0.165];
- For *implementation* this value was 0.148 [0.092, 0.207]; and
- For *advocacy* this value was 0.124 [0.063, 0.188].

It can be seen that in all three networks there was a positive relationship between *vp* and in-degree centrality, i.e. a higher popularity results in a higher likelihood of being picked elsewhere.

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 across the four networks to be determined. Extended from this, a *flexibility score* has been developed.

Given the four "pick-vectors" for each individual across 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 flexibility 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
29	0.9417932	1.2959719
24	0.9166667	1.1938123
44	0.9166667	1.1938123
39	0.8888889	1.0808731
53	0.8773830	1.0340922
9	0.8642618	0.9807440
42	0.8611111	0.9679339
52	0.8333333	0.8549948
46	0.8253796	0.8226563
32	0.8153450	0.7818577
13	0.8012695	0.7246294
49	0.7777778	0.6291165
8	0.7587129	0.5516023
38	0.7531457	0.5289673
30	0.7500000	0.5161773
37	0.7500000	0.5161773
6	0.7383582	0.4688438
50	0.7311882	0.4396920
43	0.7304439	0.4366658
31	0.7222222	0.4032382
40	0.7208776	0.3977712
11	0.6944444	0.2902990
36	0.6944444	0.2902990
51	0.6944444	0.2902990
55	0.6944444	0.2902990
34	0.6666667	0.1773599
48	0.6388889	0.0644207
25	0.5833333	-0.1614576
10	0.5833333	-0.1614576
1	0.5672227	-0.2269604
2	0.4762109	-0.5969969
45	0.4502121	-0.7027033
3	0.4446237	-0.7254246

ID	Flexibility	Z
12	0.4444444	-0.7261534
17	0.4444444	-0.7261534
14	0.4404962	-0.7422061
16	0.3928054	-0.9361080
28	0.3528374	-1.0986100
23	0.3333333	-1.1779100
35	0.0833333	-2.1943624
56	0.0833333	-2.1943624
54	0.0527864	-2.3185604
15	0.0000000	-2.5331799
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 flexibility 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 flexibility score for choosing to host a party using two different invitation methods;
- one additional point on the flexibility score for choosing to use a mix of invitation methods weighted in favour of their preferred method; and
- two additional points on the flexibility 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	0.8549948
Implementation	30	0.7500000	0.5161773
Advocacy	42	0.8611111	0.9679339

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:

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

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

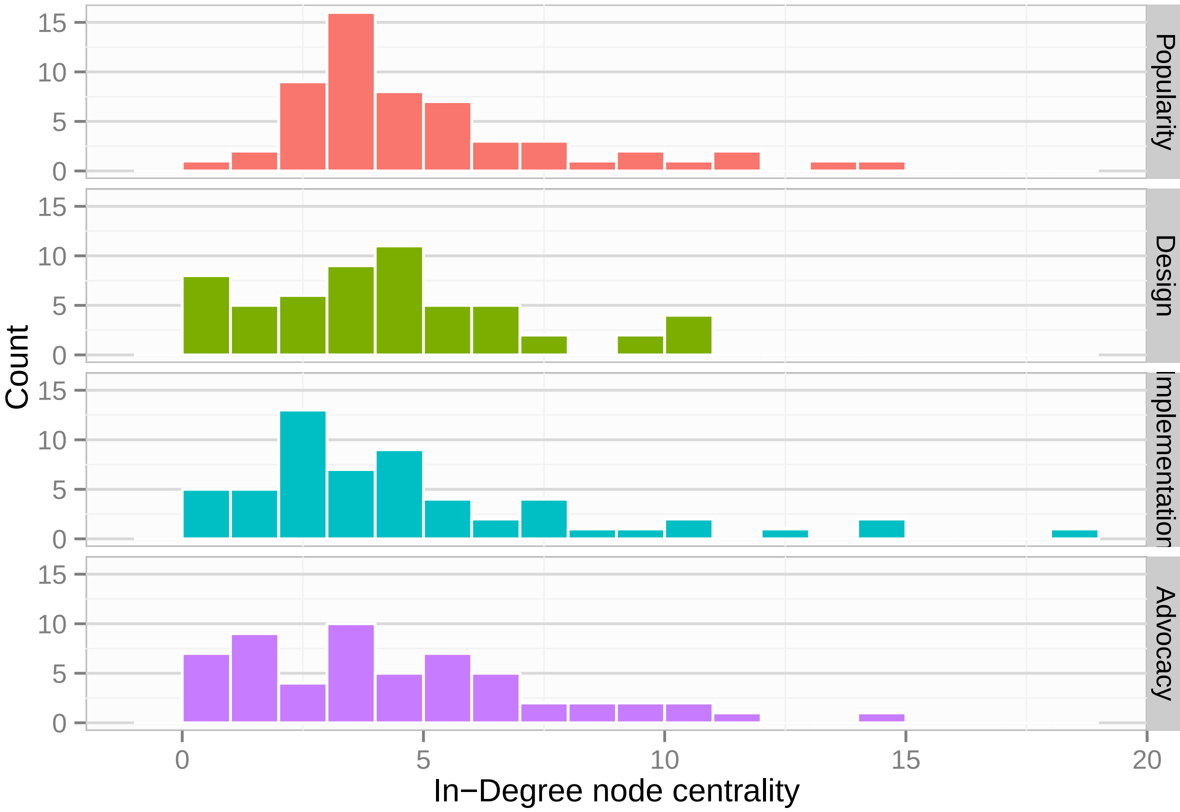
Table 8: Advocacy network measure of value

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



Appendix Two - OLS Linear Regression Results

Table 9: Regression coefficients from linear regression of design picks on visibility points (vp)

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	1.5274752	0.5857468	2.607740	0.0117105
vp	0.4939533	0.1081183	4.568636	0.0000282

Table 10: Regression coefficients from linear regression of implementation picks on visibility points (vp)

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.7660891	0.7385958	1.037224	0.3041704
vp	0.7696252	0.1363315	5.645247	0.0000006

Table 11: Regression coefficients from linear regression of advocacy picks on visibility points (vp)

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	1.4047030	0.6652786	2.111451	0.0392919
vp	0.5595827	0.1227985	4.556919	0.0000293