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

Negative binomial regression is most suitable for the problem as the data are count data and have variances greater than their means (Appendix 1). The models were created and the regression coefficients found to be:

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

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* (visibility points).

In all three models there was a positive and statistically significant (at the 1% level) relationship between visibility points and the in-degrees centrality in the other three networks. I.e. increased popularity resulted in an increased likelihood of being picked elsewhere.

(OLS regression was also performed - Appendix 2).

2 - Cosine Similarity and flexibility

The cosine similarities for each ID's picks were calculated.

The mean value for each ID was calculated, giving a single-value flexiblity score.

A higher cosine *similarity* represents a lower flexibility. Therefore the similarity was subtracted from one to give a score where a higher value indicates more flexibility. I.e. 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
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

ID	Flexibility	Z
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, leaders were picked. Qualitative information was used to enhance the flexibility score.

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

- one additional flexibility point if they chose to host a party using different invitation methods;
- one additional flexiblity point if they chose to use a weighted mix of invitation methods; and
- two additional flexiblity points if they chose a 50/50 split mix of invitation methodss.

The following three IDs were selected to be the lead member of each team. They are the individuals with the highest flexibility in each pre-picked team.

Table 5: Team leaders based on flexibility scores

Team	ID	Adjusted Flexiblity Score	Original 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 visibility points: the in-degrees centrality on the Albert Hall network) the design, implementation and advocacy scores were used. These scores are defined as:

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

Note that to make these three network properties comparable, they were mean-centered and standard-deviation-scaled. As such the combined score values range from small positive values to small negative values.

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

Table 6: Design network measure of value

ID	Network Score	Visibility Points	Measure of Value (cost-benefit)
52	1.7755496	3	0.5918499
35	1.0475553	2	0.5237777
43	0.5742077	2	0.2871039
55	1.6894286	7	0.2413469
31	1.0766333	5	0.2153267
42	1.9715924	10	0.1971592
23	0.5160858	3	0.1720286
8	0.6560915	4	0.1640229
48	1.9293307	13	0.1484101
14	1.3567138	11	0.1233376
54	0.4381155	4	0.1095289
30	0.4355138	4	0.1088784
17	0.4132363	4	0.1033091
3	0.8123977	9	0.0902664
11	0.2635357	3	0.0878452
16	0.2616838	3	0.0872279
10	0.0731026	1	0.0731026
15	0.2891459	4	0.0722865
51	1.0096977	14	0.0721213
37	0.1881685	3	0.0627228
9	0.2947380	5	0.0589476
25	0.4675078	8	0.0584385
38	0.1625666	3	0.0541889
6	0.5343061	11	0.0485733
44	0.0618393	5	0.0123679
46	0.0478870	6	0.0079812
50	-0.0859923	5	-0.0171985
1	-0.1661973	6	-0.0276996
40	-0.2502555	9	-0.0278062
2	-0.0875537	3	-0.0291846
18	-0.3861899	6	-0.0643650
13	-0.2037350	3	-0.0679117
36	-0.3300528	4	-0.0825132
29	-0.6780142	7	-0.0968592

ID	Network Score	Visibility Points	Measure of Value (cost-benefit)
47	-0.4971155	5	-0.0994231
12	-0.3055114	3	-0.1018371
57	-0.3283438	3	-0.1094479
22	-0.7786326	7	-0.1112332
33	-0.5733699	5	-0.1146740
4	-0.3181462	2	-0.1590731
41	-0.7613631	4	-0.1903408
20	-0.3853815	2	-0.1926908
24	-0.6200245	3	-0.2066748
7	-0.8089732	3	-0.2696577
39	-1.1344284	4	-0.2836071
27	-0.8659419	3	-0.2886473
56	-0.9578287	3	-0.3192762
28	-0.6425362	2	-0.3212681
21	-0.6511765	2	-0.3255882
34	-1.1344284	3	-0.3781428
49	-1.1344284	3	-0.3781428
32	-1.1344284	2	-0.5672142
45	-1.1344284	2	-0.5672142
53	-1.1344284	2	-0.5672142
26	-0.8577249	0	-Inf
5	NA	1	NA
19	NA	5	NA

Table 7: Implementation network measure of value

ID	Network Score	Visibility Points	Measure of Value (cost-benefit)
35	0.9383463	2	0.4691731
43	0.5114222	2	0.2557111
55	1.3229148	7	0.1889878
23	0.5563988	3	0.1854663
57	0.5367165	3	0.1789055
18	0.9619756	6	0.1603293
30	0.5803550	4	0.1450888
11	0.4011297	3	0.1337099
42	1.1960673	10	0.1196067
48	1.5141486	13	0.1164730
34	0.3077572	3	0.1025857
52	0.2990011	3	0.0996670
2	0.2964968	3	0.0988323
39	0.3554677	4	0.0888669
25	0.6509160	8	0.0813645
3	0.7233702	9	0.0803745
21	0.1364949	2	0.0682475
7	0.2022918	3	0.0674306
6	0.7340910	11	0.0667355
4	0.1316244	2	0.0658122
51	0.7748445	14	0.0553460
46	0.3272241	6	0.0545374
13	0.1623150	3	0.0541050
50	0.2296320	5	0.0459264

ID	Network Score	Visibility Points	Measure of Value (cost-benefit)
29	0.2674192	7	0.0382027
1	0.2093629	6	0.0348938
22	0.2014916	7	0.0287845
9	0.1412127	5	0.0282425
27	0.0833702	3	0.0277901
44	0.1295870	5	0.0259174
45	0.0510116	2	0.0255058
38	0.0759528	3	0.0253176
47	0.0953711	5	0.0190742
40	0.0993263	9	0.0110363
14	0.0395235	11	0.0035930
31	-0.0205126	5	-0.0041025
15	-0.0326986	4	-0.0081747
41	-0.1076543	4	-0.0269136
20	-0.0639984	2	-0.0319992
10	-0.0456409	1	-0.0456409
33	-0.2341201	5	-0.0468240
8	-0.2151790	4	-0.0537948
37	-0.2008014	3	-0.0669338
49	-0.2311631	3	-0.0770544
36	-0.3258818	4	-0.0814705
54	-0.3965701	4	-0.0991425
24	-0.3362952	3	-0.1120984
16	-0.6262556	3	-0.2087519
28	-0.5382289	2	-0.2691145
17	-1.9682485	4	-0.4920621
12	-1.9437790	3	-0.6479263
56	-1.9970034	3	-0.6656678
53	-1.9665925	2	-0.9832963
32	-1.9970034	2	-0.9985017
26	-1.9970034	0	-Inf
5	NA	1	NA
19	NA	5	NA

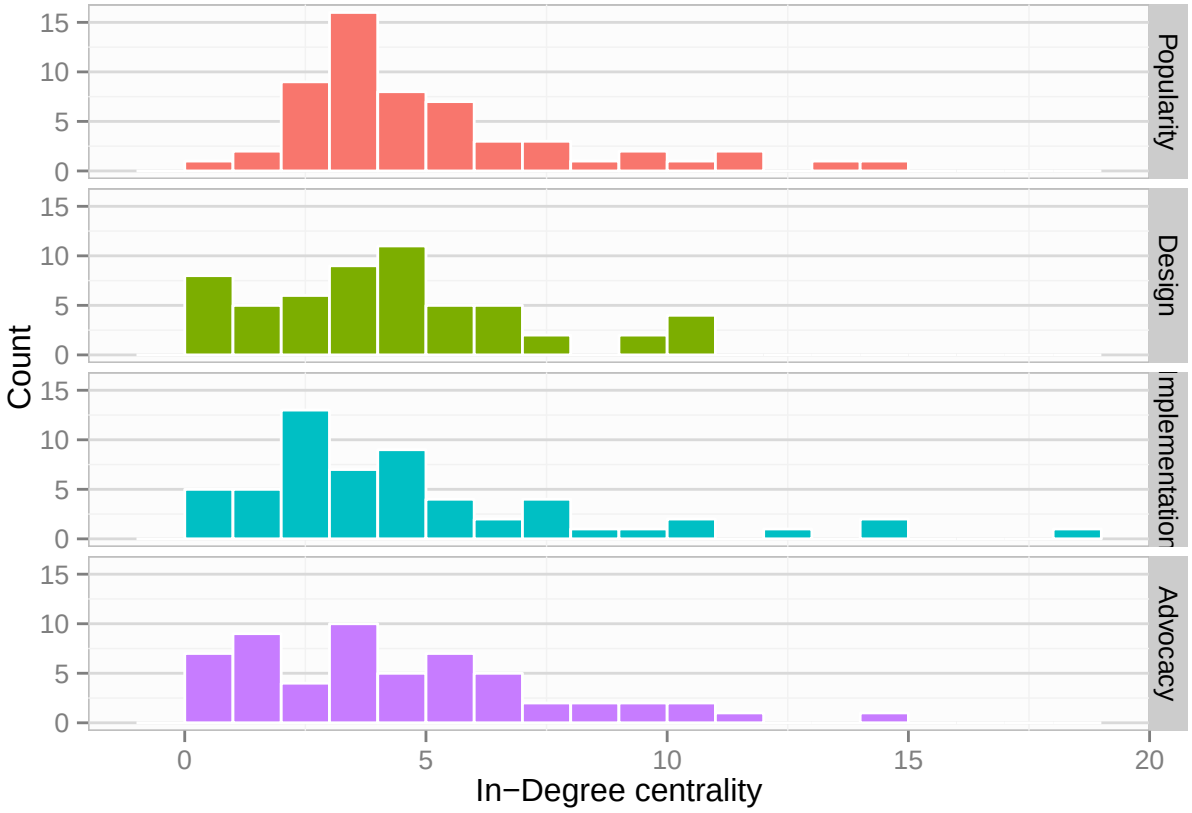
Table 8: Advocacy network measure of value

ID	Network Score	Visibility Points	Measure of Value (cost-benefit)
35	0.9404589	2	0.4702294
10	0.4688547	1	0.4688547
50	1.5437607	5	0.3087521
25	2.1487924	8	0.2685991
54	0.9118415	4	0.2279604
52	0.5710575	3	0.1903525
23	0.5405137	3	0.1801712
30	0.7121953	4	0.1780488
42	1.6906367	10	0.1690637
57	0.5009658	3	0.1669886
18	0.7400452	6	0.1233409
55	0.7906126	7	0.1129447
48	1.2908717	13	0.0992978
6	0.9554678	11	0.0868607

ID	Network Score	Visibility Points	Measure of Value (cost-benefit)
56	0.2285706	3	0.0761902
40	0.5834349	9	0.0648261
1	0.3727096	6	0.0621183
4	0.1055407	2	0.0527703
9	0.2385837	5	0.0477167
3	0.2719974	9	0.0302219
46	0.1245581	6	0.0207597
51	0.2641880	14	0.0188706
14	-0.0017206	11	-0.0001564
29	-0.1315220	7	-0.0187889
11	-0.1131409	3	-0.0377136
2	-0.1287779	3	-0.0429260
47	-0.2332233	5	-0.0466447
39	-0.2628181	4	-0.0657045
22	-0.4676482	7	-0.0668069
37	-0.2137078	3	-0.0712359
15	-0.2850678	4	-0.0712669
17	-0.2972158	4	-0.0743039
8	-0.3233655	4	-0.0808414
44	-0.4699151	5	-0.0939830
33	-0.5178695	5	-0.1035739
12	-0.3227137	3	-0.1075712
31	-0.5595238	5	-0.1119048
24	-0.3545014	3	-0.1181671
38	-0.3780787	3	-0.1260262
43	-0.2786073	2	-0.1393036
19	-0.7246687	5	-0.1449337
41	-0.6537524	4	-0.1634381
28	-0.3868711	2	-0.1934356
16	-0.5906137	3	-0.1968712
36	-0.7959320	4	-0.1989830
34	-0.6109640	3	-0.2036547
49	-0.6128073	3	-0.2042691
7	-0.6206129	3	-0.2068710
32	-0.4570769	2	-0.2285385
13	-1.4798265	3	-0.4932755
45	-1.4798265	2	-0.7399133
53	-1.4798265	2	-0.7399133
26	-0.7634617	0	-Inf
5	NA	1	NA
20	NA	2	NA
21	NA	2	NA
27	NA	3	NA

Appendices

Appendix One - Exploratory plot of in-degree centrality distributions



As the data are count date, and the conditional variances of the distributions were much larger than the conditional means, negative binomial regression was chosen as the most suitable method for this regression problem.

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

Linear models all reveal a statistically significant, positive relationship between visibility points and picks elsewhere. These models were assessed with data of size 57 and have R^2 values of 0.275, 0.367, 0.274 respectively.