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].

- 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.

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
16	0.4522774	-0.6987900

ID	Flexibility	Z
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

Appendices

Appendix One - Exploratory plot of in-degree centrality distributions