Identifying Sociological Trends in facebook Networks

Amanda L. Traud ■ *Department of Mathematics* ■ *Carolina Population Center Trainee* ■ *The University of North Carolina at Chapel Hill*

Introduction

- Networks are simply a set of objects, or nodes, that are connected in some way. They can consist of objects of any kind from chemicals to people. The networks studied consisted of users on Facebook.com and their "friends".
- Facebook is a social networking website that plays a prominent role in college life. The data for this project was taken from Facebook in 2005. There are currently over thirty million members. Facebook sets up a network for each college or university; five of these networks were examined closely.
- From those five networks, algorithms were used to detect communities, or tightly connected sets of users. Then these communities were correlated with characteristics given by those users.

Methods

- Community Detection
  - Why is it important?
    - Breaking networks in to communities can help us understand the structure of the network.
    - Community Detection can be used on many types of networks: biological networks, disease networks, and terrorist networks, just to name a few.
  - Methods
    - Spectral Graph Partitioning – Will not work because of the size of our networks.
- Newman's Leading Eigenvector Method
  - In this method, one tries to maximize a quantity called modularity, which is the actual number of connections in a community minus the expected number.
  - To do this:
    - Put all data into an adjacency matrix like the one below.

Names	Kelly	John	Michael	Kristen
Kelly	0	1	1	0
John	1	0	1	0
Michael	1	1	0	1
Kristen	0	0	1	0

    - Calculate the modularity matrix
    - Take the leading eigenvector of the modularity matrix
    - From the entries in the leading eigenvector, put objects into two groups.
    - Repeat process on each subsequent group until modularity of the network is maximized.
    - The formulas for this method are described above and to the right.

Formulas for Newman's Leading Eigenvector Method

- Modularity: $Q = \frac{1}{4m} s^T B s$
- Modularity Matrix – $B = A - \frac{k_i k_j}{2m}$
- Placement Vector
  - 1 if place in Leading Eigenvector is positive
  - -1 if place in Leading Eigenvector is negative

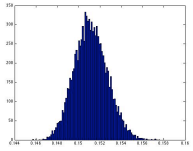
Q	Modularity
A	Adjacency Matrix
B	Modularity Matrix
m	Total Connections or sum of k
k	Degree Vector or sum over A
s	Placement Vector

- Similarity Coefficients
  - To compare communities to given characteristics, one calculates similarity coefficients.
  - These coefficients are based on the idea that the characteristics given are a new set of communities.
  - For the five similarity coefficients we calculated, we had to pair every node with every other node and count pairs. All five of the Similarity coefficients calculated are based on different combinations of the quantities a, b, c, and d.
    - a = number of pairs of nodes in the same community in the first set of communities, and in the second set.
    - b = number of pairs of nodes in the same community in the first set, but not in the same community in the second set.
    - c = number of pairs of nodes in the same community in the second set, but not in the first.
    - d = number of pairs of nodes in different communities in both sets.

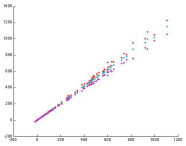
Formulas for Similarity Scores

- Jaccard = $\frac{a}{a+b+c}$
- Rand = $\frac{a+d}{a+b+c+d}$
- Folkes-Mallows = $\frac{a}{\sqrt{(a+b)(a+c)}}$
- Minkowski = $\sqrt{\frac{b+c}{b+a}}$
- Gamma = $\frac{(a+b+c+d)a-(a+b)(a+c)}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$

- Z-Score Calculation
  - After calculating similarity coefficients, it was not apparent what value of similarity score would give the best fit between communities and characteristics.
  - Permutation tests were performed on the similarity coefficients to create a distribution of coefficients
  - Many other tests were performed, including calculating the kurtosis and making histograms of the distributions to make sure the distributions were Gaussian. (See below for a distribution example)



- After finding the distribution was Gaussian, the Z-score, or the number of standard deviations better than the mean, was calculated of the similarity coefficient, of the original communities and characteristics ,compared to the distribution.
- It was found that no matter what similarity coefficient we used, the Z-score was approximately the same for each category, as shown by the plot below.



Graph of Z-scores versus Average Z-score for each category
Note: The slope is approximately 1

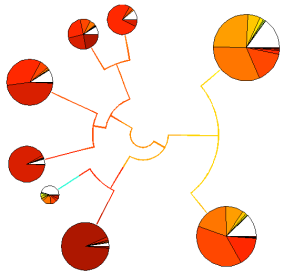
Princeton Similarity Coefficients

Characteristic	Folkes-Mallows	Gamma	Jaccard	Minkowski	Rand
Major	0.2185	0.0817	0.1066	1.0476	0.7234
House	0.2004	0.0257	0.1046	1.1055	0.692
Year	0.3994	0.2692	0.2341	0.9726	0.7616
High School	0.0964	-0.0108	0.0355	1.0461	0.7241

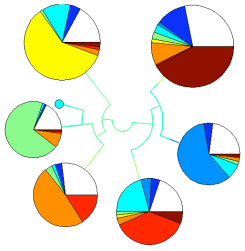
Average of the Z-scores for the five similarity coefficients

Name	Major	House	Year	High School
Princeton	43.739	9.908	442.58	-4.5876
Georgetown	1.9642	114.44	653.7	23.3684
UNC	23.283	123.33	598.87	17.0125
Caltech	3.5362	133.1	7.5456	6.1529
Oklahoma	7.514	25.667	9.6241	21.7821

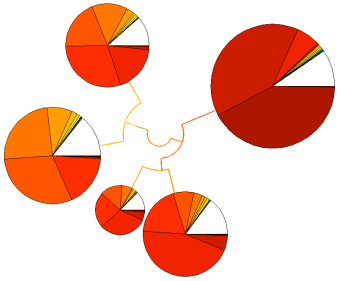
Results



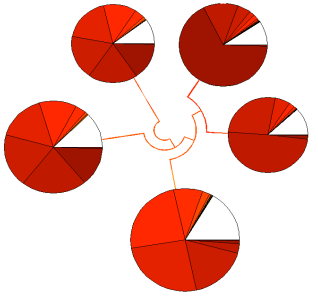
Georgetown: Colored by graduation year



California Technical Institute: Colored by house



Princeton: Colored by graduation year



UNC – Chapel Hill: Colored by graduation year

Conclusion/Contact info

- While a number of different similarity coefficients appear in the literature, the statistics of those as obtained through permutation tests were approximately the same, specifically the Z-scores.
- Most schools break up into communities due to graduation year; however, some ,like Caltech, break up into communities by house, which appears to be consistent with the social structure of this college.
- These analyses were performed both ignoring and respecting the fact that there is missing data. In almost no case did the differences change the qualitative conclusion.
- Contact: Amanda L. Traud (altraud@email.unc.edu)

Acknowledgements

- My Collaborators: Eric Kelsic, Peter J. Mucha, Mason A. Porter
- UNC-Chapel Hill AGEP Program, Director: Dr. Valerie Ashby
- National Science Foundation HRD-0450099, DMS-0645369

UNC ALLIANCE FOR GRADUATE EDUCATION & THE PROFESSORIATE

UNC CAROLINA POPULATION CENTER