Sociological Trends in facebook Networks

Amanda L. Traud
Advisers: Dr. Peter J. Mucha and Dr. Mason A. Porter
Mathematics Department
Carolina Population Center
The University of North Carolina at Chapel Hill

Introduction

- Networks are 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 September 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

- Many different community detection algorithms exist, we used Newman's Leading Eigenvector method which works well for large networks.
  - This method maximizes modularity which is a measure of the number of intracommunity edges relative to that expected on average.
  - We then have an optimization problem that we're solving by computing the leading eigenvector of a "modularity matrix".
- We compared community structure to demographic traits of individuals.
  - We calculated similarity coefficients based on pair counting of every single possible pair in the network, checking to see if the nodes in each pair are in the same community and have the same demographic characteristic. (a=# of pairs in same community and with same demographic trait, b=# of pairs in same community with different demographic traits, c=# of pairs with same demographic traits and in different communities, and d=# of pairs in different communities and with different demographic traits)
  - After calculating these coefficients, it was hard to tell what was a "good" value, so we used permutation tests to calculate the Z-score of each coefficient for each trait.

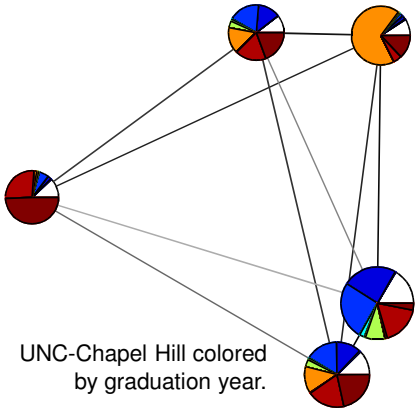
Formulas for Similarity Scores

- 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 = $\frac{(a+b+c+d)a-(a+b)(a+c)}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$

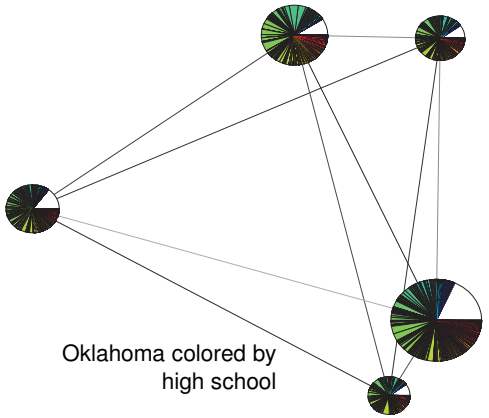
Conclusion/Contact info

- 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. The Z-scores for Rand, Folkes-Mallows and Gamma are identical and a formula is available for this value.
- These analyses were performed using three different protocols to deal with missing data, all missing demographic traits were put in their own category, just the nodes with missing traits for each category were removed, and then all nodes with any missing demographic traits were removed. In almost no case did the differences change the qualitative conclusion.
- Contact: Amanda L. Traud (altraud@email.unc.edu)

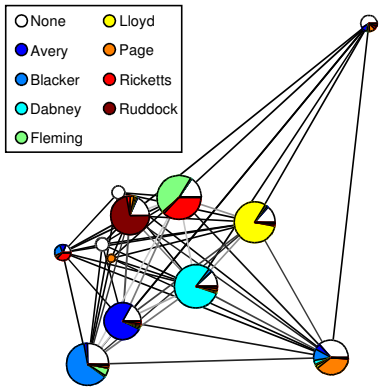
School	1st Dominating Trait	2nd Dominating Trait
Caltech	House	None
UNC	Graduation Year	Dorm
Oklahoma	Dorm	High School
Princeton	Graduation Year	Dorm
Georgetown	Graduation Year	Dorm



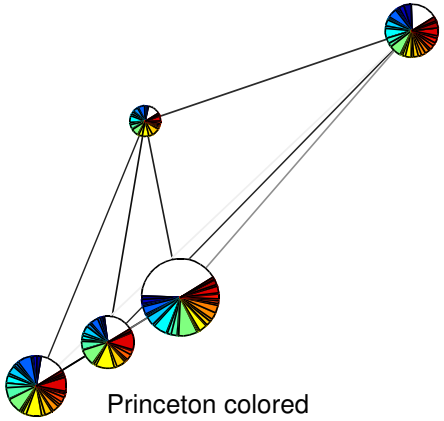
UNC-Chapel Hill colored by graduation year.



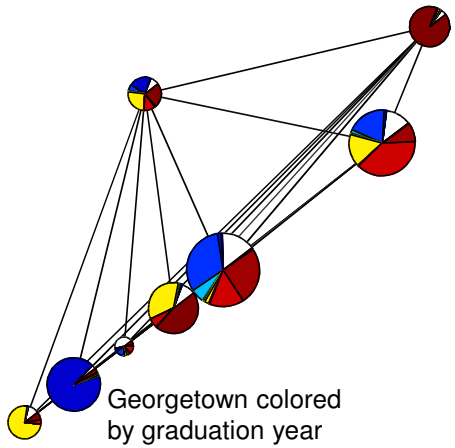
Oklahoma colored by high school



Caltech breaks up into twelve communities, most of which are dominated by one particular house as shown above.



Princeton colored by major



Georgetown colored by graduation year

Acknowledgements
My Collaborator: Eric Kelsic
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