Emergent (Non)Majors: Communities and Connections in an Interdisciplinary College

Samuel Heller

and

David P. Feldman

College of the Atlantic

2 June 2008

`http://hornacek.coa.edu/dave/`

Outline

1. Motivation (Dave)
2. Introduction to Networks, Definitions of Key Terms (Dave)
3. Methods and Results (Sam)
4. Discussion and Observations (Sam)
5. Conclusion (Dave)
6. Discussion (Everybody)

- Please ask clarifying questions during the talk.
- To the extent possible, we will try to save larger questions for the discussion period after the presentation.

Motivation: Are there Patterns in Students' Course Choices?

- College of the Atlantic has essentially no course requirements beyond introductory and breadth requirements.
- Students thus have broad latitude in choosing courses.
- How do students exercise this latitude? Are there any central tendencies or trends? Are there different groupings or cliques of students with similar choices?

Two Possible Views on Student Course Choices and Curricular Structure

1. Students create original, interdisciplinary areas of concentration.
   - Students follow their own paths and do not choose narrow foci
   - Students at COA thus combine areas of study in ways that are not common at other college
   - However, students may thus leave COA without gaining depth or expertise in any area
2. COA has *de facto* majors
   - Students cluster into distinct groupings or cohorts
   - Students take a narrow set of classes, re-creating the disciplinary majors that the college seeks to avoid
   - Some faculty offer classes of interest only to this narrow set of de-facto majors.

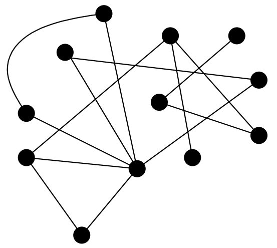
One College or Many?

1. Do all students design original majors, and hence we are One College, united by our individuality?
2. Do students (and faculty) segregate themselves along disciplinary lines?
3. Or, perhaps students (and faculty) form communities, but along interdisciplinary lines?

- There are tools in the emerging area of network analysis that make possible an empirical examination of these questions.
- In order to explain these tools, we need to begin with some basic ideas and definitions from the field of complex networks.

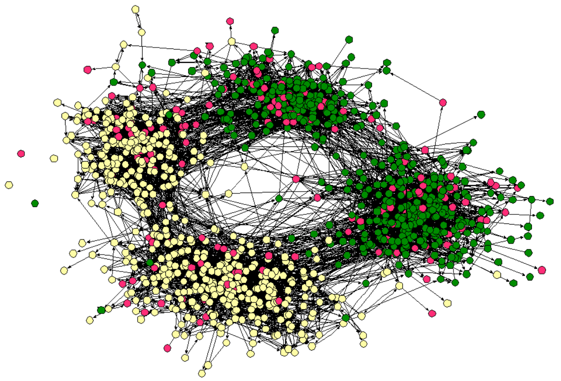
What is a Network?

1. A collection of **nodes**
2. A collection of **edges** connecting nodes



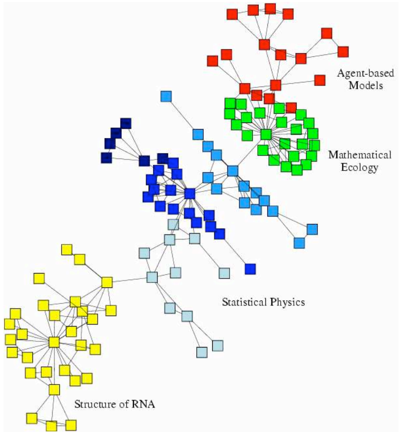
- A network model treats all nodes and links the same
- In a picture of a network, the spatial location of nodes is arbitrary
- Networks are abstractions of connection and relation
- Networks have been used to model a vast array of phenomena

Network Example 1: High School Friendships



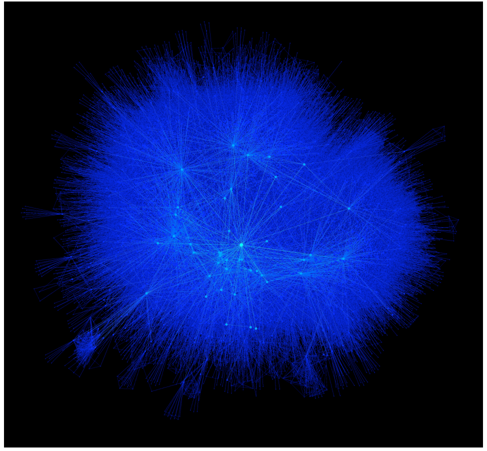
- Nodes = Students, Links = Friendships, Color = Race
- Data: J. Moody, Race, school integration, and friendship segregation in America, *American Journal of Sociology* 107, 679-716 (2001).
- Figure: M.E.J. Newman, The structure and function of complex networks, *SIAM Review* 45, 167-256 (2003). `www-personal.umich.edu/~mejn/networks/`

Network Example 2: Interdisciplinary Collaborations



- Nodes = Researchers, Links indicate that the researchers have co-authored one or more papers.
- Figure: M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 8271-8276 (2002).

Network Example 3: Online Social Network



- Nodes = Accounts (47,471) on Friendster, Links (432,420) indicate that accounts are friends.
- Figure: Jeffrey Heer. `http://www.cs.berkeley.edu/~jheer/socialnet/`

Network Properties

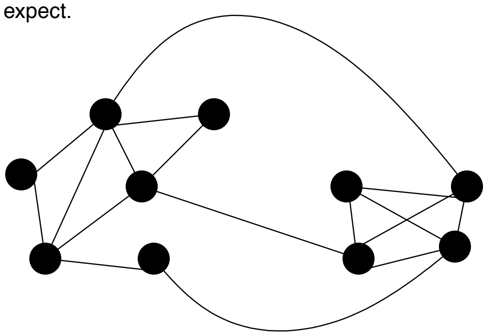
Given a network, there are a number of structural questions we may ask:

1. How many connections does the average node have?
2. Are some nodes more connected than others?
3. Are there clusters or groupings within which the connections are particularly strong?

- We will focus on this last question
- A group of nodes which are connected more strongly to each other than to the rest of the network is called a **community**.

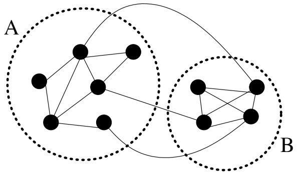
What Makes a Community?

- Suppose we suspect that a network is made of two communities. Can we test this?
- A group is a community if there are more within-community connections than one would expect.



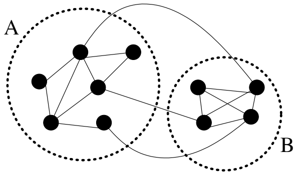
- How can we quantify this?

Modularity: A Measure of Community-ness



- Suppose we think there are two communities, A and B.
- Divide the links into two types: between-community and within-community.
- For this network, there are 8 links within A, 6 within B, and 3 between A and B.
- There are 17 total links.
- So $\frac{8}{17}$ of the links are within community A.
- Is this a lot? How many would we expect?

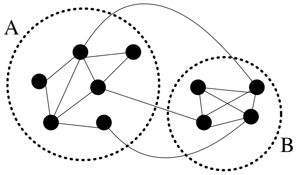
Modularity: Continued



- 8 links within A, 6 within B, and 3 between A and B, and 17 total links.
- $\frac{8}{17}$ of the links are within community A. Is this a lot?
- Of the 17 total links, 11 connect to A.
- If no community structure, then the communities edges link to are independent.
- So, if we draw a link at random, what is the chance it connects A to A?

$$\text{Prob of connection to A} \times \text{Prob of connecting to A} = \frac{11}{17} \times \frac{11}{17}$$

Modularity: Conclusion

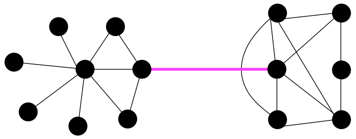


- **Modularity** is defined as the fraction of within-community links minus the number of within community links one would expect if the links were random.
- For community A: $\frac{8}{17} - \frac{11^2}{17^2}$.
- For community B: $\frac{6}{17} - \frac{9^2}{17^2}$.
- Adding these together, we get the modularity of the network. In this case, modularity $= 0.12$.
- **Modularity is a measure of the strength of a set of communities. The bigger the number, the stronger the community structure.**

Discovering Communities?

- Modularity tells us how to test the strength of a set of communities.
- But, how can we discover communities?
- Sam will describe an algorithm for this task.
- We want our algorithm to find a set of communities with a large modularity.
- We don't want to specify beforehand how many communities to look for.
- But first, one more network property:

Betweenness



- The betweenness is a property of an edge.
- Betweenness measures how important an edge is in connecting other members of the network.
- To calculate betweenness, consider all possible pairs of edges.
- Find the shortest path connecting each pair.
- The betweenness of an edge is the number of shortest paths running along that edge
- Idea: Edges with high betweenness separate communities.
- See, e.g., Finding and evaluating community structure in networks, M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004). for discussion of betweenness and modularity.

Algorithm

Introduction

Data

Network

Processing

Community Discovery

- Girvan-Newman Algorithm
- General weighted network algorithm
- Detects by removing specific ‘weak’ links
- Provides a set of community structures and each structures’ associated modularity

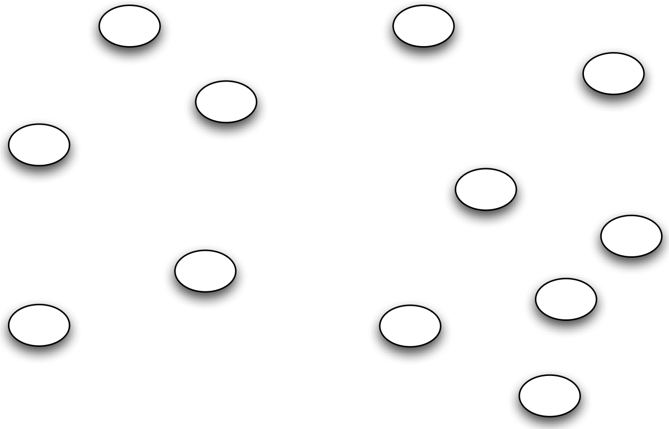
Data

- 4 graduating classes (2004, 2005, 2006, 2007)
  - – Students graduated under 5 years
  - – 147 total students
- Removed Independent studies, senior projects
- Removed classes of less than 5 total students

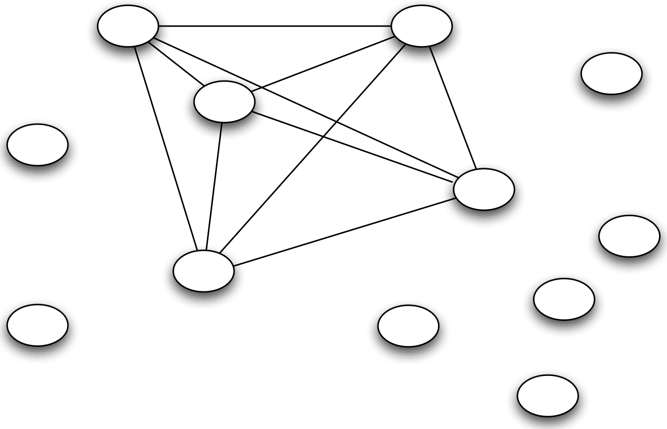
Network Generation

- Nodes are courses
- Link: all courses a student has taken to all other courses taken by student
  - – Link 1/geometric mean $(C_1+C_2)/(C_1*C_2)$
  - – ^ Purpose: two small classes have strong link while 2 large classes have weak link (makes algorithm independent of course size)
- Repeat for all students
- Remove: all edges with weight less than threshold
  - – ^Purpose: remove extremely weak links to increase speed and effectiveness of algorithm

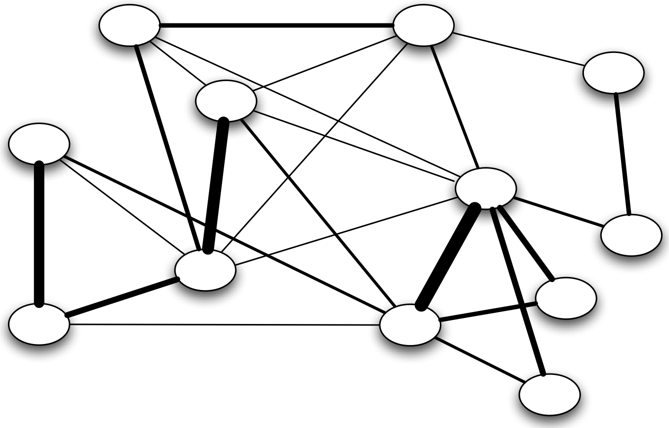
Empty Set of Courses



One Student Added



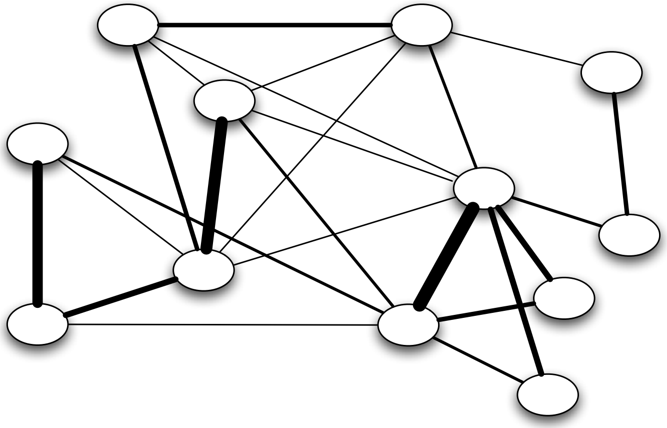
All Students



Create Communities

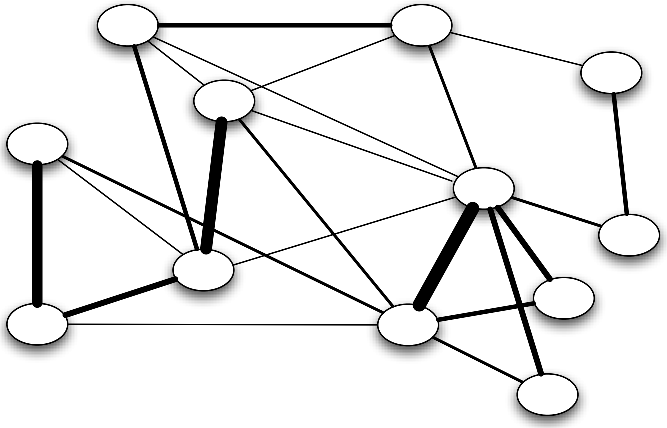
- Input: Network created above
- Determine and remove most between link
- Repeat until all links are removed
- Result: sequence of candidate community structures
- Calculate modularity for each candidate
- Choose structure with highest modularity

Output from Network Generation



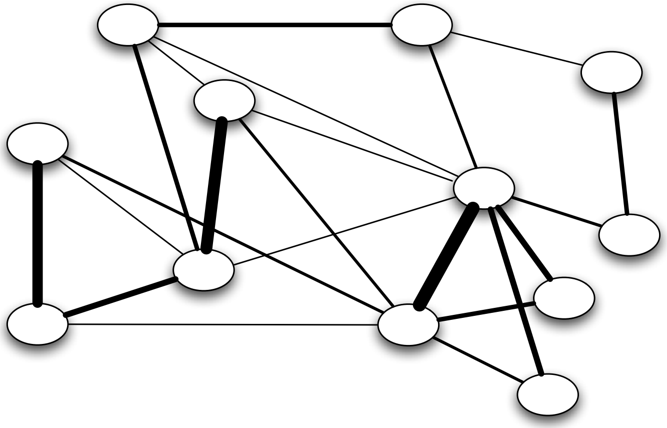
Modularity: 0.0

Remove most between link



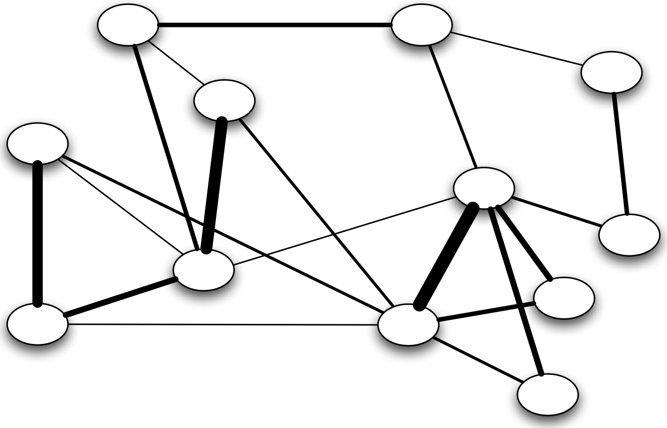
Modularity: 0.0

Remove most between link



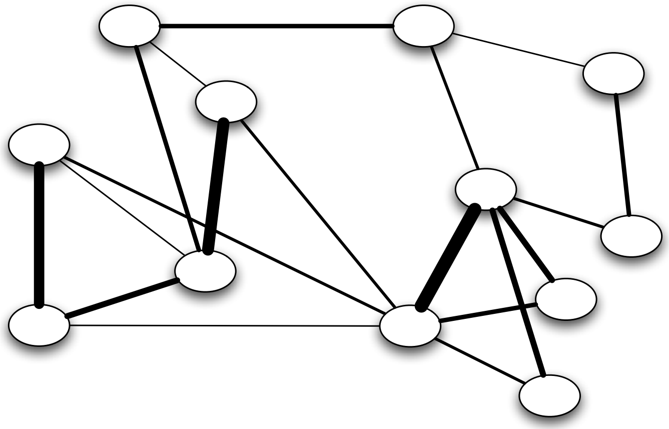
Modularity: 0.0

Remove most between link



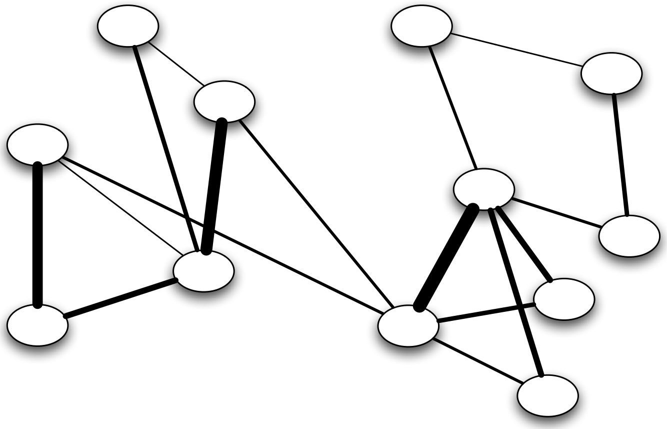
Modularity: 0.0

Remove most between link



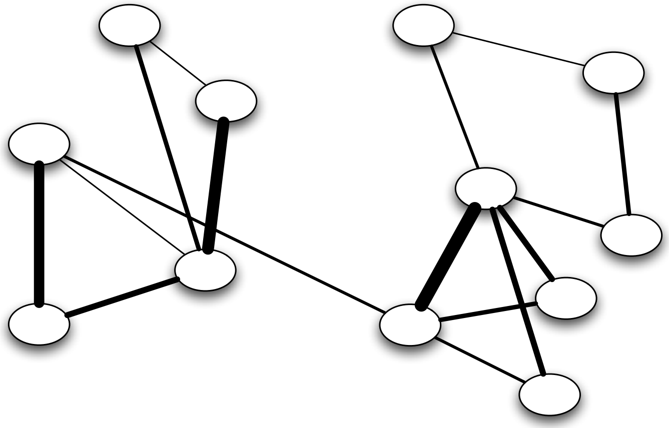
Modularity: 0.0

Remove most between link



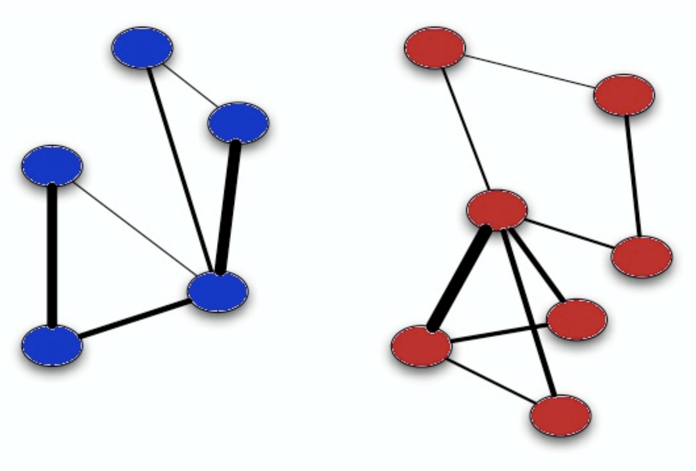
Modularity: 0.0

Remove most between link



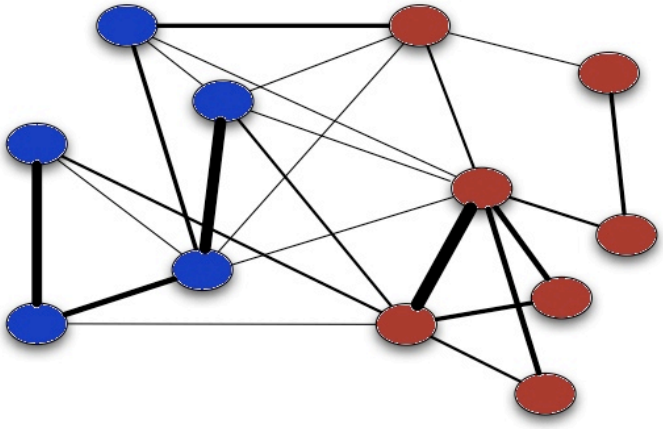
Modularity: 0.0

Maximum modularity – 2 communities



Modularity: 0.20

Compare Original with Algorithm



Effectiveness of community structure? **Modularity**

Control – Random students

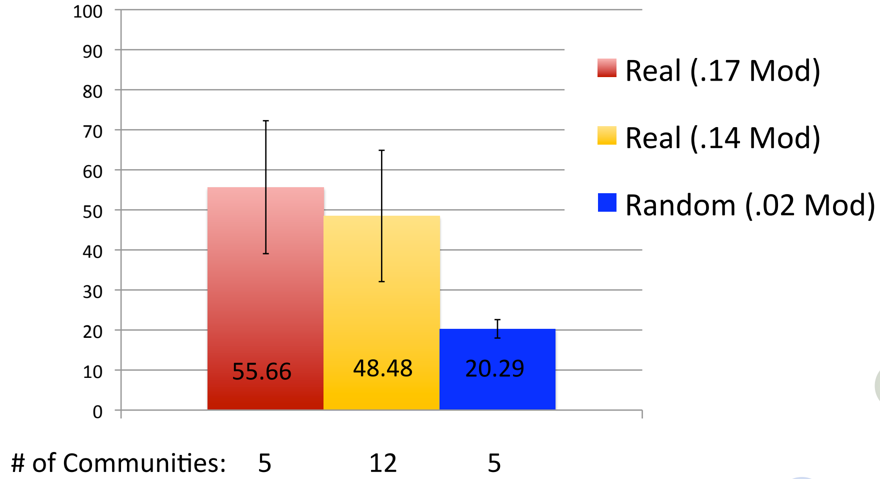
- Created random students
  - Random students have number of courses that fit distribution of real students (μ = 29, σ = 4)
  - Random students take courses with probability proportional to course size
  - Random students fulfill requirements (2xES, 1xHY..)

Results

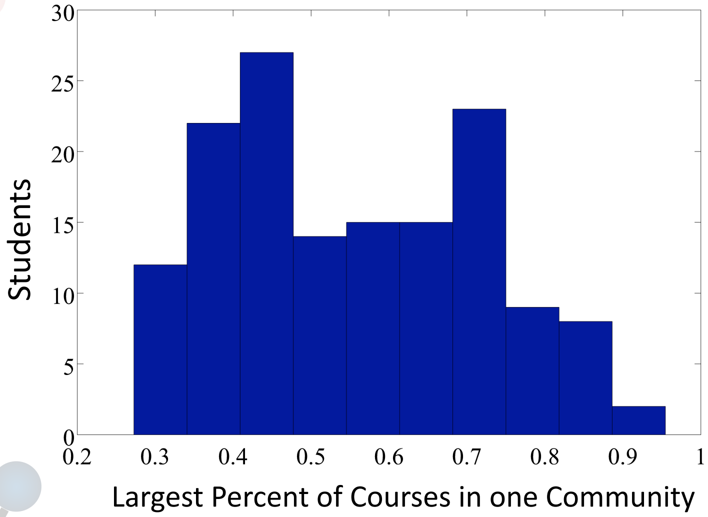
General

- The community structure with highest modularity has:
  - 5 communities
  - modularity = 0.1748
- Students do not take courses at random
- Average student takes roughly 55% of classes *considered* in one community

Average percentage of courses students take in one community



Histogram: Percentage of courses in one community



Community Data

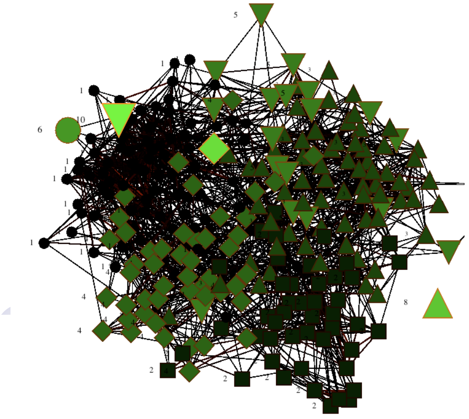
Community Number	Overall	1	2	3	4	5
Courses	259	76	47	67	51	18
Credits taken	2781	950	491	736	449	155
Students in community (highest percent of courses in one community)	147	61	22	46	15	4
Average percent in community for above students	56%	60%	53%	52%	56%	44%
Profs in community (highest percent of courses in one community)	36	13	7	8	6	2
Average percent in community for above profs	71%	87%	60%	81%	82%	58%

Results

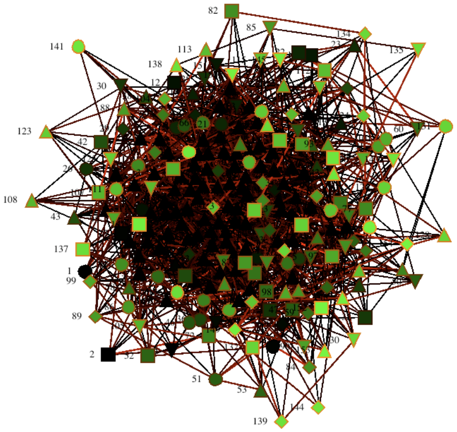
- Real (.17 Mod) statistically significant difference from Random (P-value = 1 x 10^{-59})
  - Students do not take courses randomly
  - Instead they take an impressive amount in specific categories (average 55% within one community)
- Given n = 147, and only 5 communities, each community fits a large number of students

Visualization shows clear clustering not seen in random network

Real (.17 Mod)

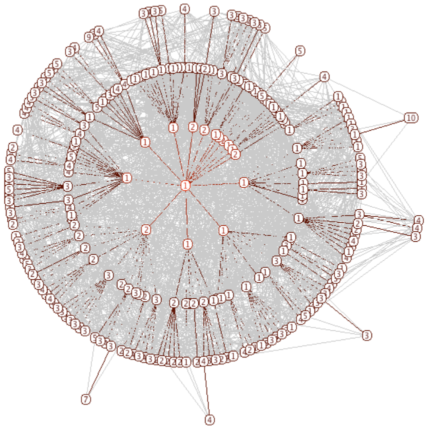


Rand (.02 Mod)

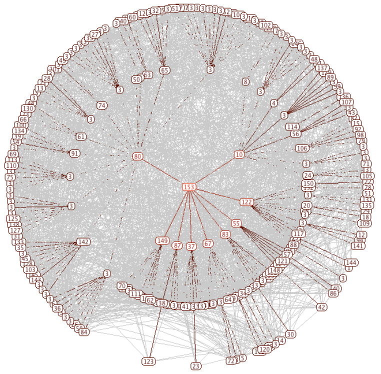


Radial Tree Graph

Real (.17 Mod)



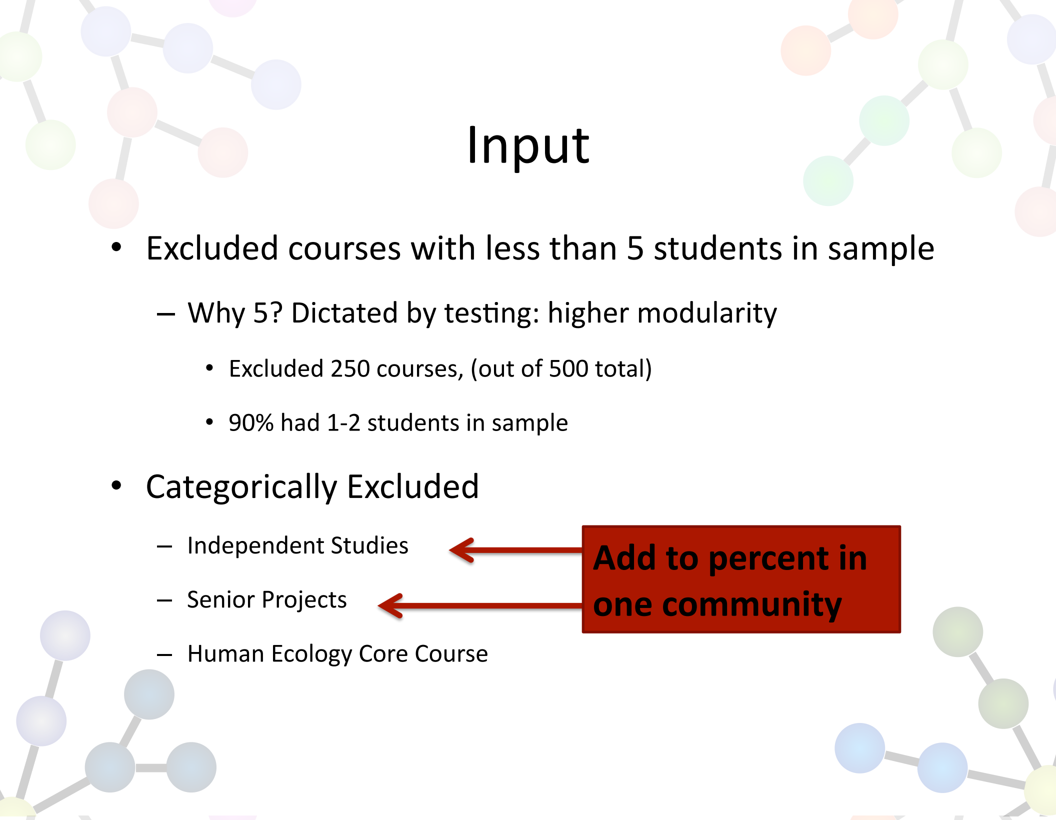
Rand (.02 Mod)



Students as Nodes

- Very low modularity: .05
- Suggests students do not take courses based on other students
- Does not provide thematic results, but only specific cases
  - Hard to look at a set of students classified as a ‘community’ and understand what it means

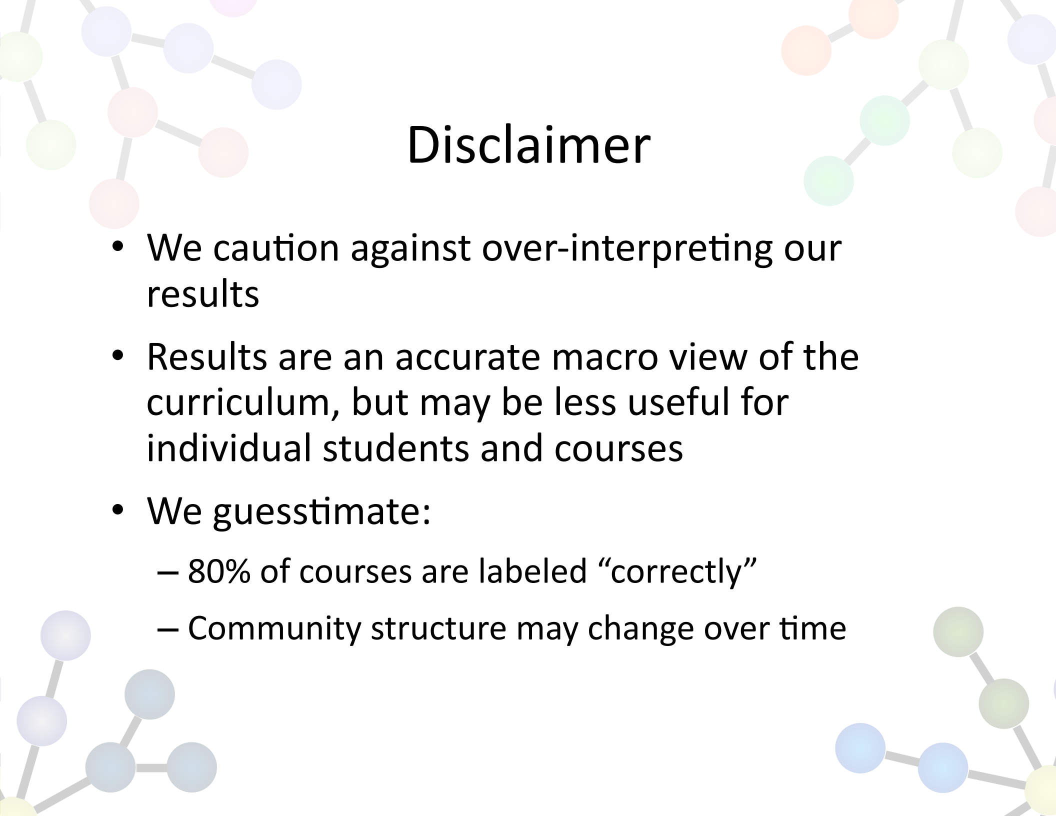
Discussion



Input

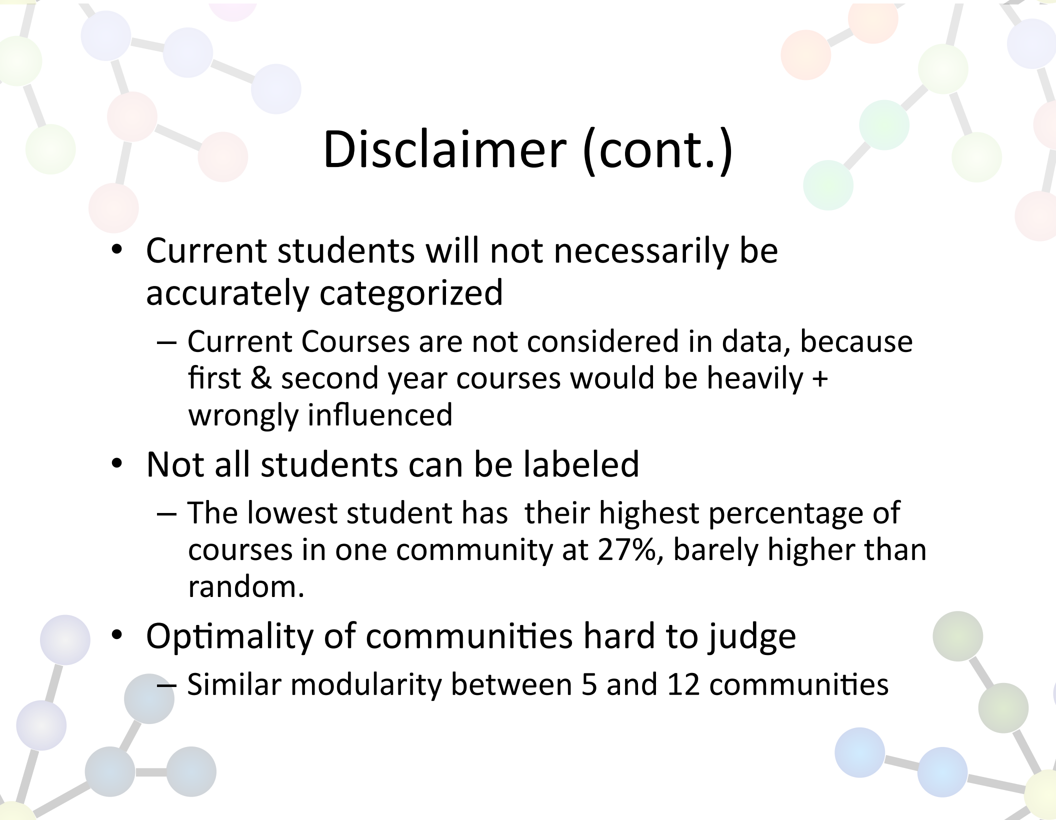
- Excluded courses with less than 5 students in sample
  - Why 5? Dictated by testing: higher modularity
    - Excluded 250 courses, (out of 500 total)
    - 90% had 1-2 students in sample
- Categorically Excluded
  - Independent Studies
  - Senior Projects
  - Human Ecology Core Course

Add to percent in one community



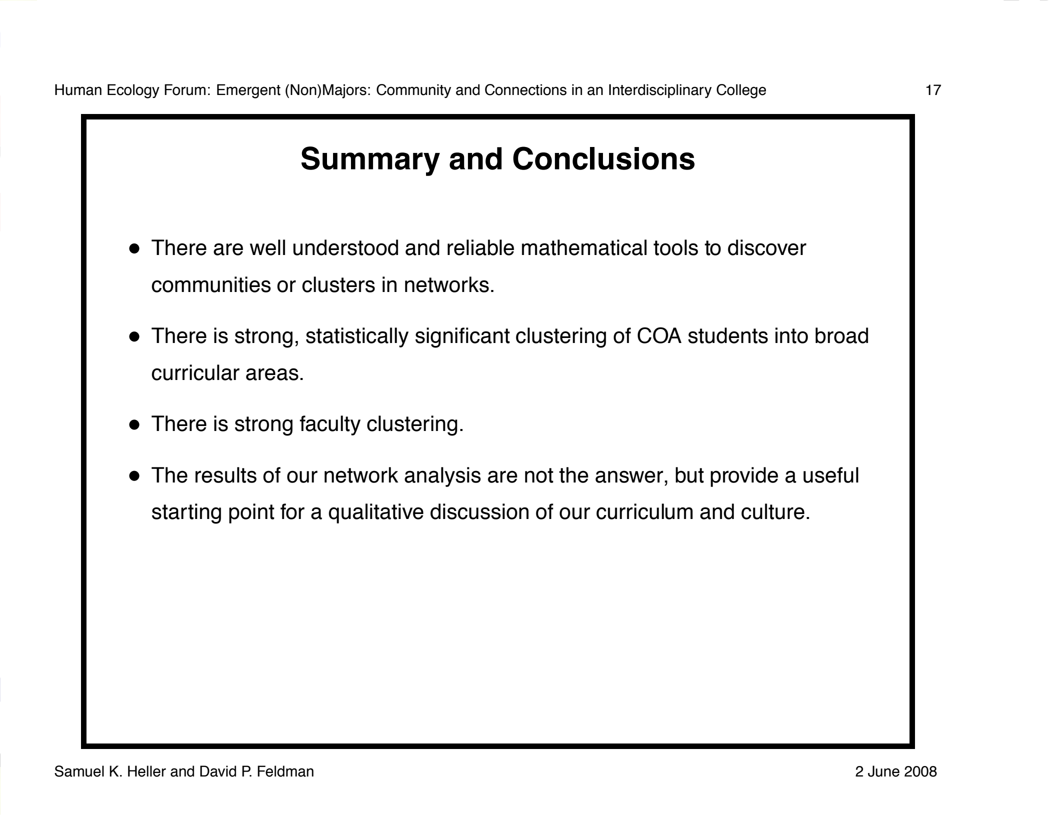
Disclaimer

- We caution against over-interpreting our results
- Results are an accurate macro view of the curriculum, but may be less useful for individual students and courses
- We guesstimate:
  - 80% of courses are labeled “correctly”
  - Community structure may change over time



Disclaimer (cont.)

- Current students will not necessarily be accurately categorized
  - Current Courses are not considered in data, because first & second year courses would be heavily + wrongly influenced
- Not all students can be labeled
  - The lowest student has their highest percentage of courses in one community at 27%, barely higher than random.
- Optimality of communities hard to judge
  - Similar modularity between 5 and 12 communities



Human Ecology Forum: Emergent (Non)Majors: Community and Connections in an Interdisciplinary College 17

Summary and Conclusions

- There are well understood and reliable mathematical tools to discover communities or clusters in networks.
- There is strong, statistically significant clustering of COA students into broad curricular areas.
- There is strong faculty clustering.
- The results of our network analysis are not the answer, but provide a useful starting point for a qualitative discussion of our curriculum and culture.

Samuel K. Heller and David P. Feldman 2 June 2008

Acknowledgments

- Judy Allen, for providing much data.
- Aaron Clauset, for advice about community detection algorithms for weighted networks.
- Allen Downey, for suggestions about bootstrapping to estimate the strength of our communities.
- Jay McNally, for advice about community detection algorithms.
- John Carver, Sarah Dreup, Nick Jenei, Sarah Drerup, Sarah Jackson, Sarah Short, Cecily Swinburne, Amy Wesolowski, for ideas about characterizing the five communities.

Future Work

- Analyze larger data sets
- Compare with other colleges, both those with majors, and those without
- Look at course choices over longer time frame to see if the community structure has changed significantly over time
- If one were to classify students' community membership after two years of courses, does it predict their community membership after four years?
- Network analysis lends itself to an analysis of the connections and communities in a variety of other settings. Students can explore this further in the networks class next fall.

Discussion Questions

In our opinion

- These clusters are about “right,” in that they indicate a healthy degree of broad concentration.
- The goals of interdisciplinarity and self-direction may be in conflict.
- If we are committed to student self-direction, we should be open to the possibility that students chose a disciplinary direction.

Some questions:

- What level of clustering would be ideal?
- What are the common features, if any, of the courses within each cluster?
- What names would you assign to the five clusters?
- What, if anything, are the broader implications of our results for COA?