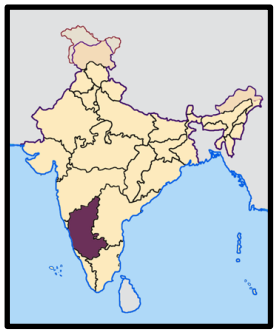
A Study of Social Networks in Karnataka, India



Lisbeth Acosta, Jamaris Burns, Emmie Román
Dr. Onnela & Patrick Staples

HARVARD
SCHOOL OF PUBLIC HEALTH

Overview

Network science can be used to learn more about the behavior of and interaction between individuals, biological pathways, and more.

In this presentation we will explain fundamental network science terminology and concepts in relation to a specific network data set retrieved from villages in Karnataka, India

- social networks
- modularity
- homophily


India
Karnataka

Karnataka Network Data

Data source: MIT Economics Department

Survey of social networks in 75 villages in rural, southern Karnataka, India

- Household census and individual level demographics obtained
- Individuals were asked questions about relationships they had with others in their village
- Example question:
  Who do you borrow money from?

Social Network Terminology

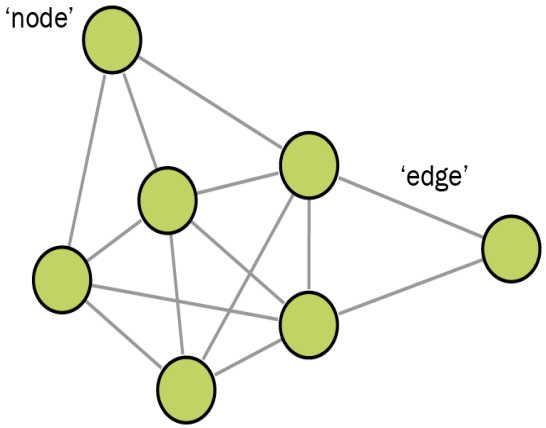
Graphing connections helps identify network structure

Karnataka data set is *social* network data

Node: One individual

Edge: Social interaction between two people (i.e. friendship)

Degree: # of connections an individual has

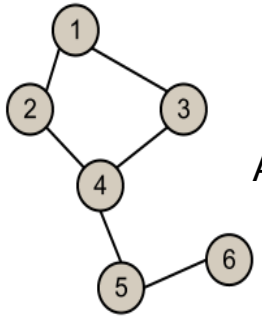


Mathematical Representation

Network data is represented by adjacency matrices, usually denoted by A

For undirected networks

A is symmetric, meaning:



Undirected Graph

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$A_{ij} = \begin{cases} 1 \text{ if there is an edge between i and j} \\ 0 \text{ otherwise} \end{cases}$$

Communities within a Network

- Within social networks, can exist communities

- Communities have a high density of connections that form tight-knit clusters within a network

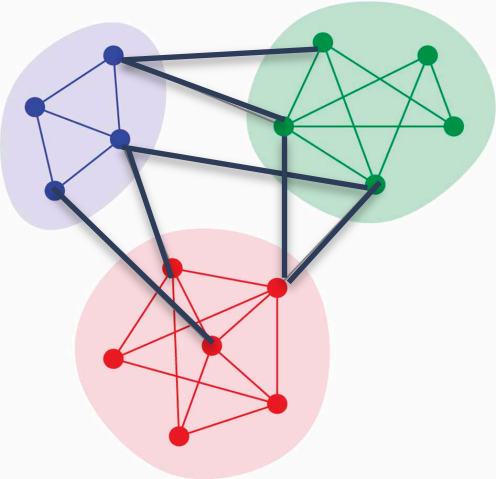
- In the data, a community could be a group of households within a village that interact more amongst themselves than they do with households outside their group

Nodes/ people can belong to more than 1 community

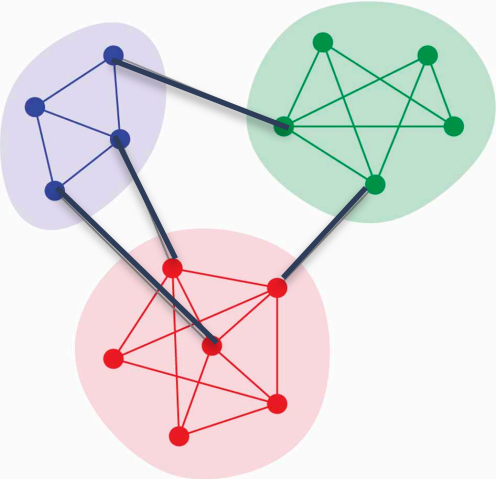
For research purposes, it is ideal to find the most defined communities

- Highest possible number of connections within clusters and least possible number of connections outside the clusters

Less Defined



More Defined

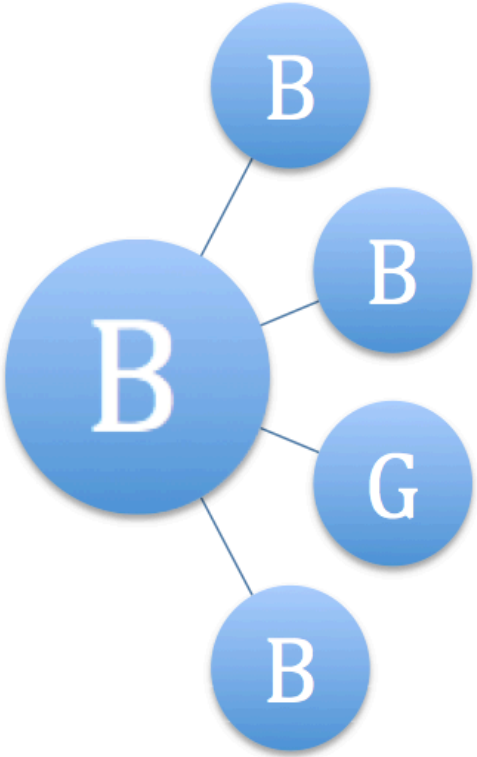


Homophily

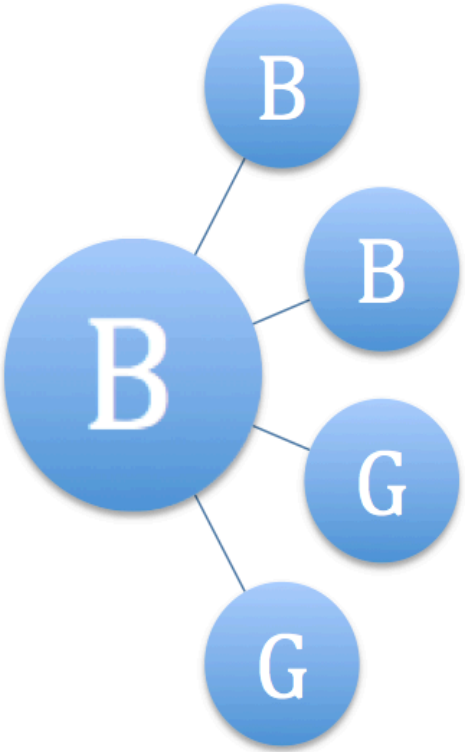
Homophily studies the basis between network ties.

Do people form more social relationships with people who exhibit similar attributes than with those who exhibit less common attributes?

Example: Do those who have unhealthy eating patterns form more relationships with those who exhibit the same eating patterns than those who have healthier eating patterns?



Observed



Expected

Question

Are the villages in the Karnataka data set homophilous?

Tested for homophily between villages on the basis of

1. Gender
2. Caste
3. Savings

Methods

Performed binomial tests in python using the Karnataka village data set

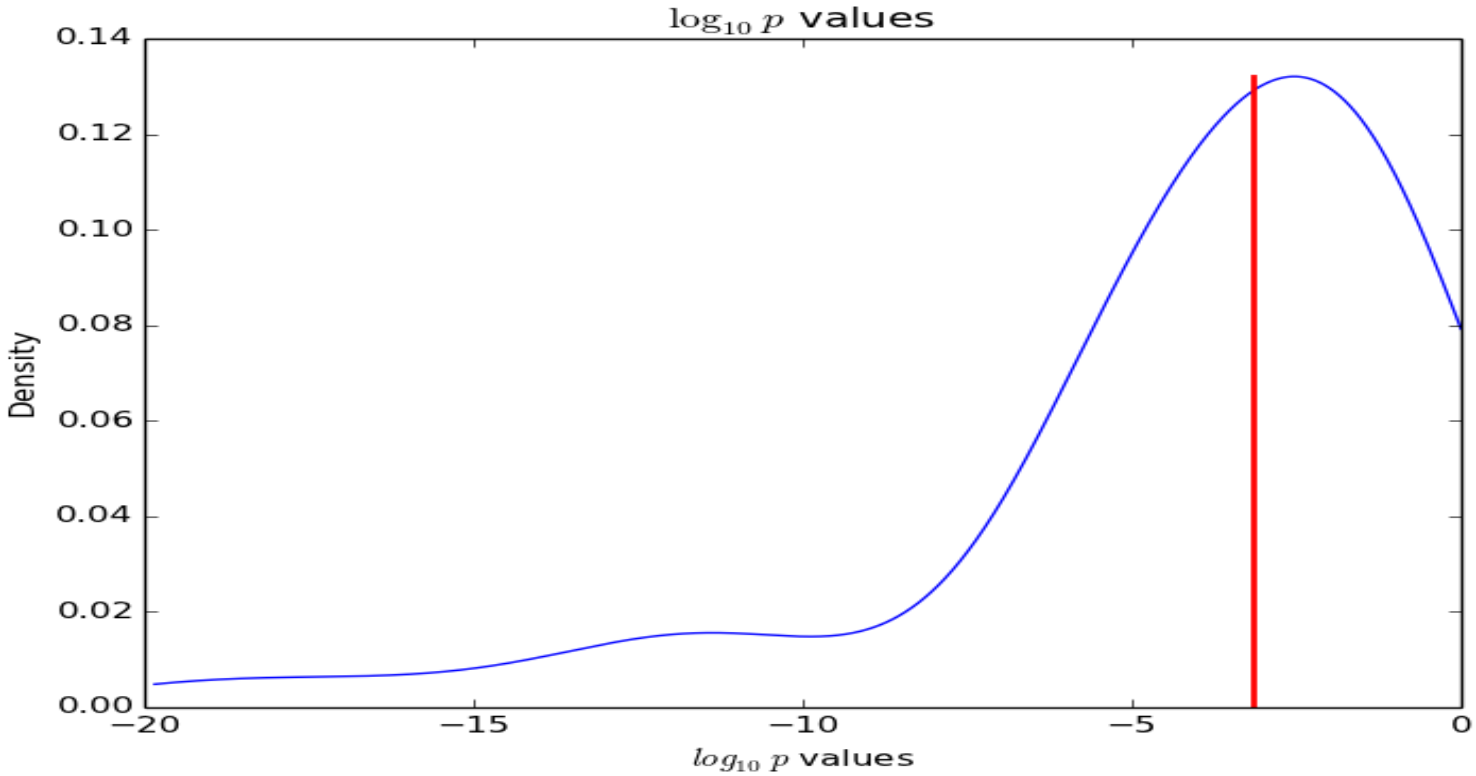
- Ran simultaneously for each village

Test resulted in a distribution of homophily p-values for each village

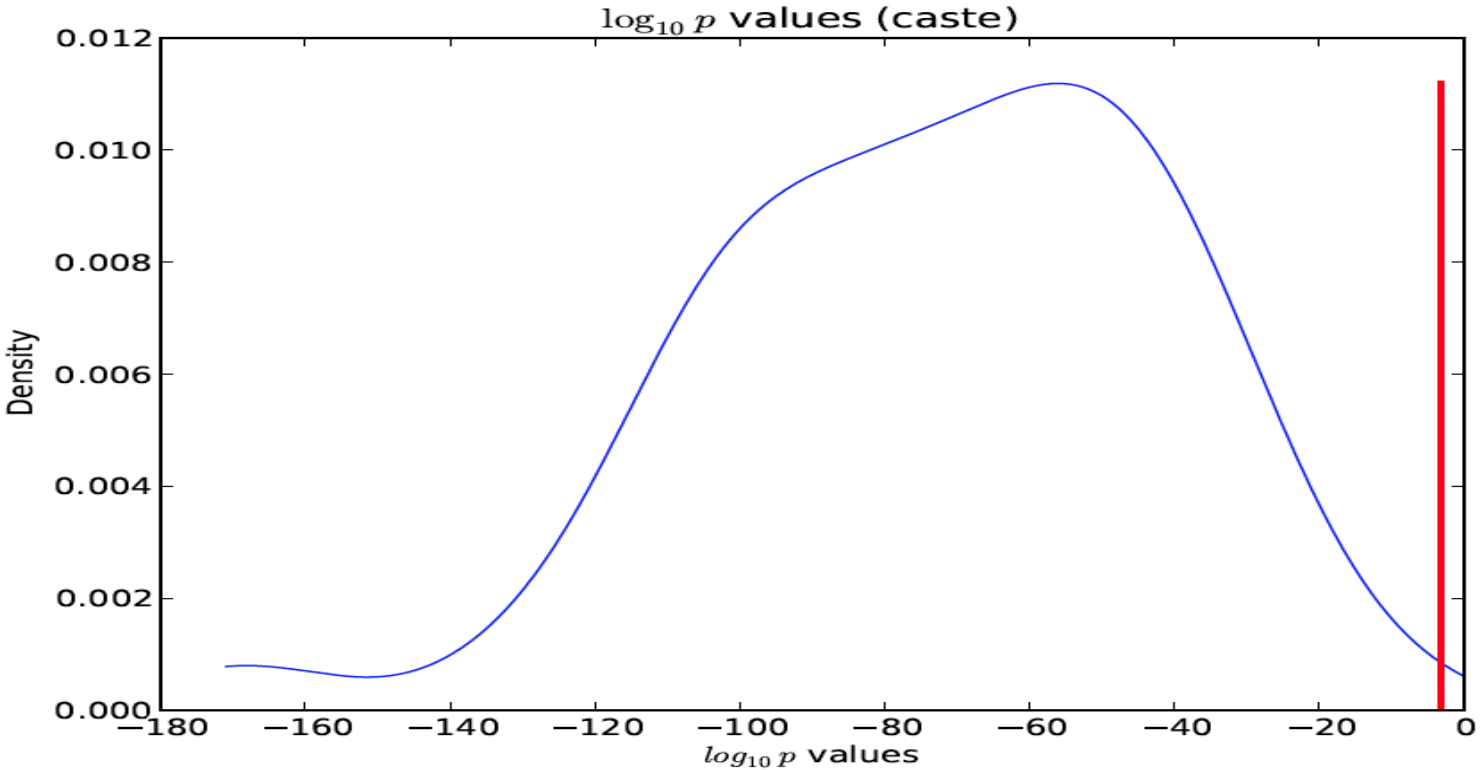
Distribution reflects multiple testing

- Some villages had significant p-values for homophily on basis of pre-determined attribute (caste, gender, or savings) and others did not
- Bonferroni Correction created an appropriate critical value to take into account invalid p-values that can appear when performing multiple testing

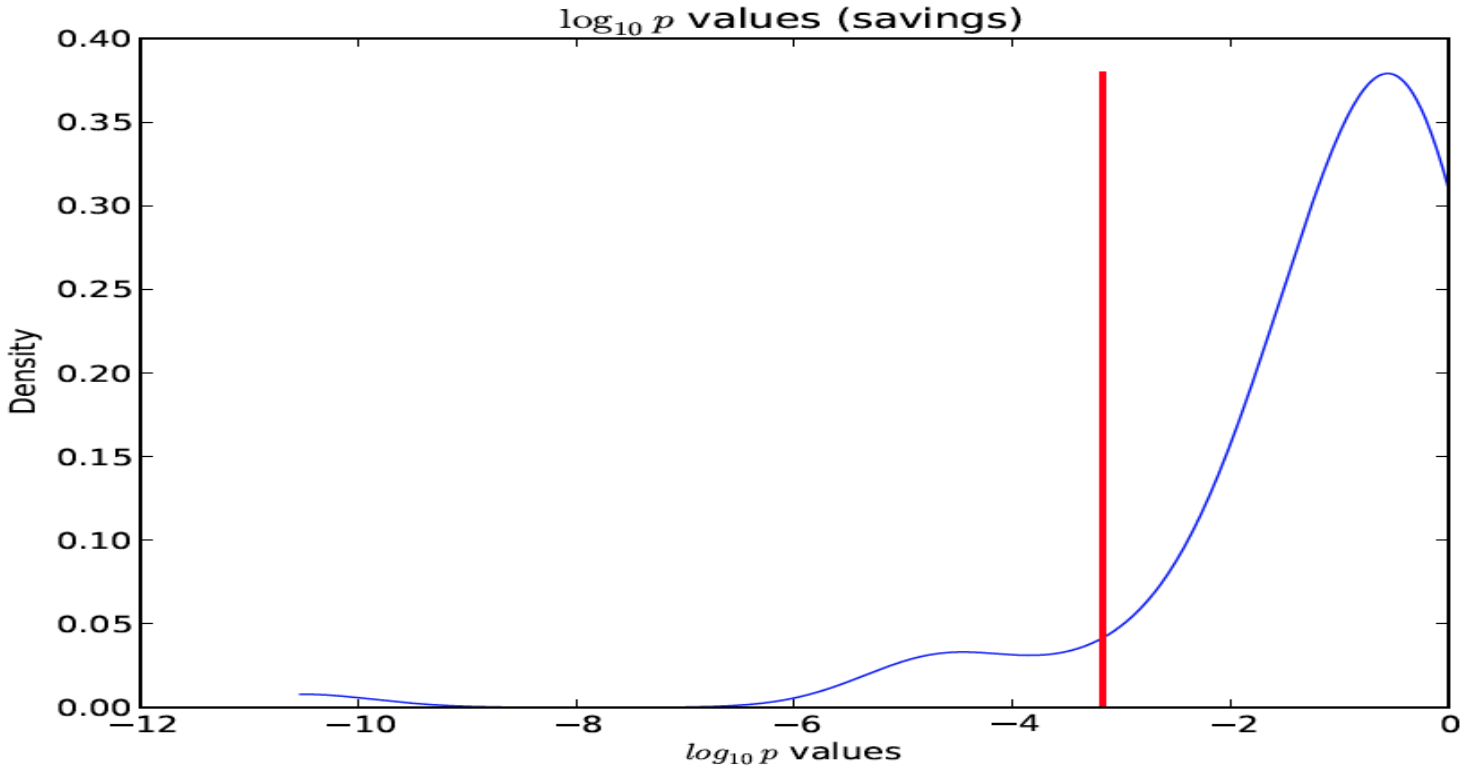
Gender


$\log_{10} p$ values
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
−20
−15
−10
−5
0
log_{10} p values

Caste


$\log_{10} p$ values (caste)
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−180
−160
−140
−120
−100
−80
−60
−40
−20
0
$log_{10} p$ values

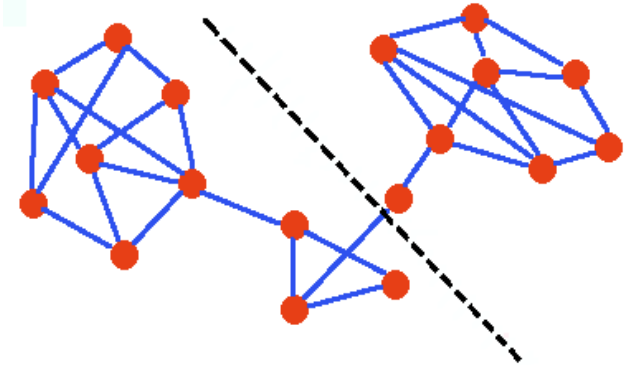
Savings


$\log_{10} p$ values (savings)
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−12
−10
−8
−6
−4
−2
0
$log_{10} p$ values

Identifying Clusters within Networks

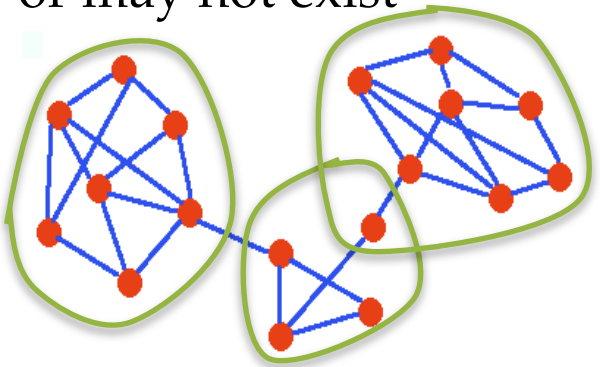
Graph Partitioning

- Nodes (people) are placed in specific communities

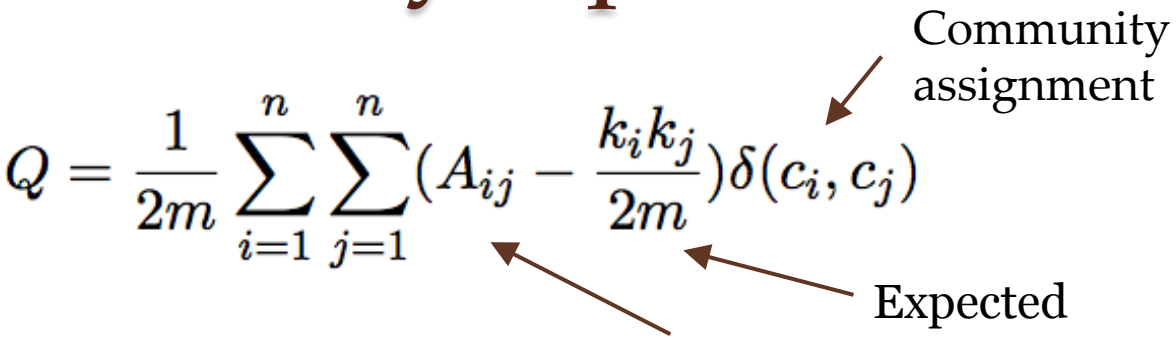


Community Detection

- Allows nodes (people) to naturally divide into communities
- Note: A community may or may not exist



Modularity equation



A_{ij}= Adjacency matrix element (0,1)

n = Number of nodes

k_i= Degree of each node

m = Number of edges

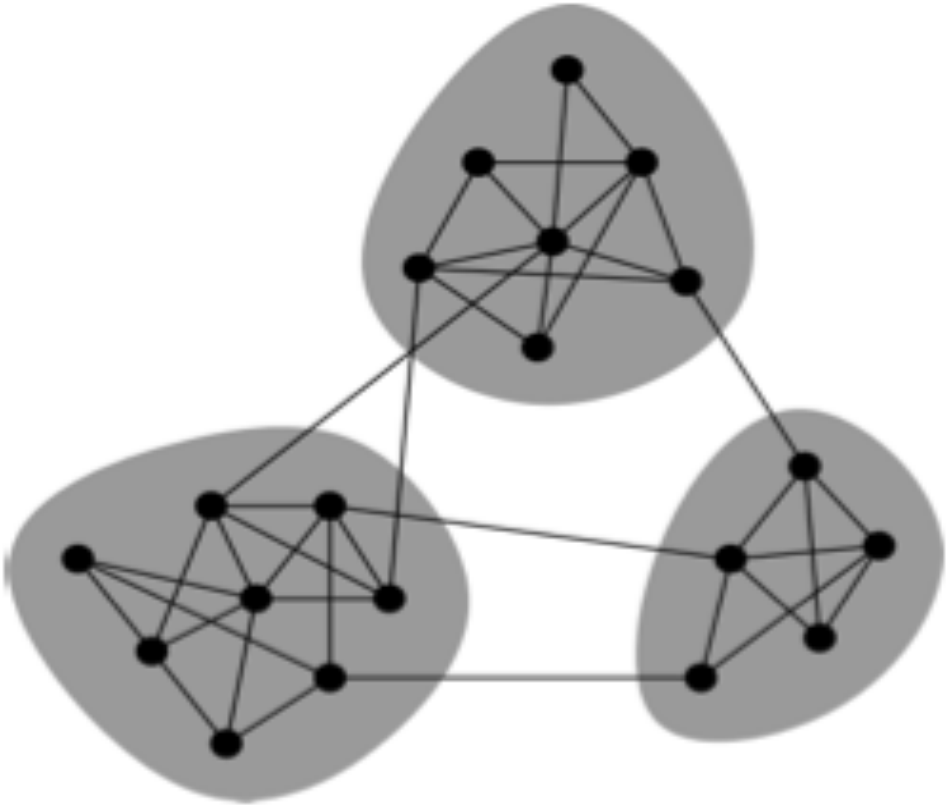
δ = Kronecker delta function (0,1)

c_i= Community assignment of each node

Maximization of Modularity

Goal: To find the best division(s) where

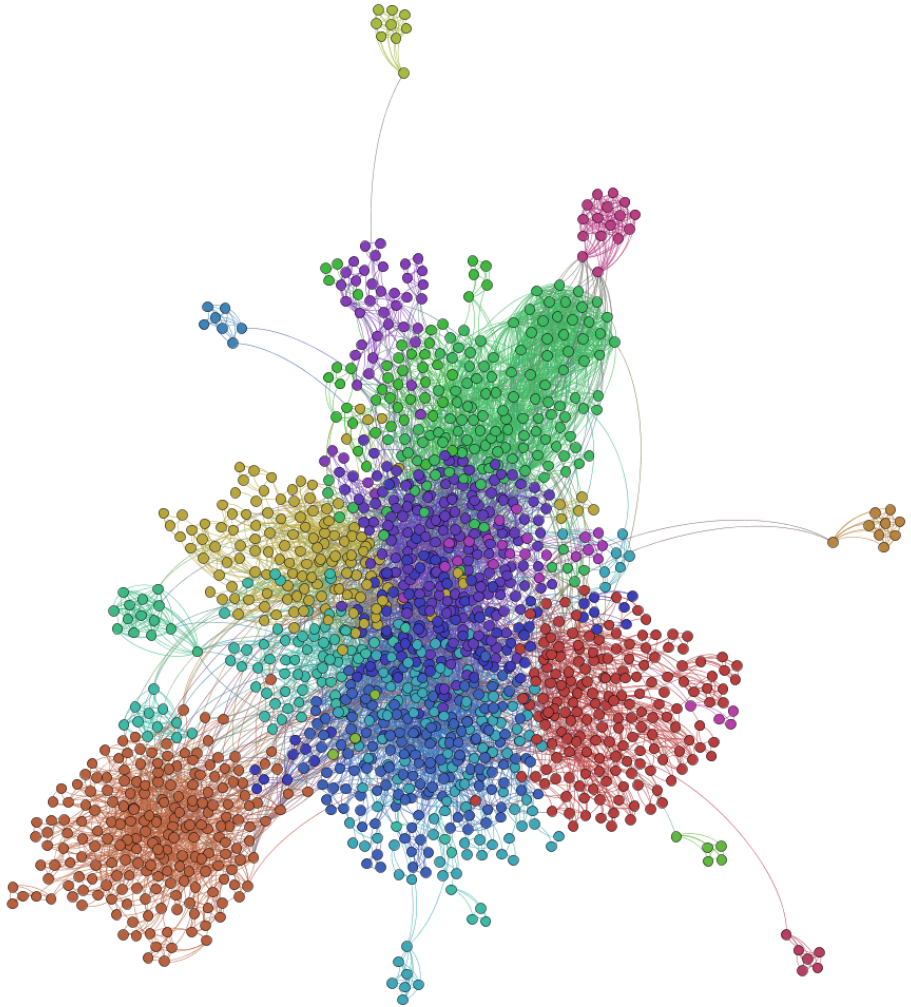
- ✓ Maximize connections within communities
- ✓ Minimize connections between communities



Newman M. E. J. (2006) *PNAS* **103**, 8578-8582

Kernighan-Lin Algorithm

- Non-trivial problem
- Starts with randomized community assignments
- Flip community assignment and run the algorithm to observe what happens to Q equation
- If the switch increases the Q value, store it
- Repeat this process with all assignments



Limitations

- Missing information from village 13 and 22
- Kernighan-Lin algorithm only splits into 2 communities
- Homophily is not defined on community structures
- Not clear how values of significance compare with that of other countries

Conclusion and Suggestions

- Our homophily findings across villages offer insight to possible homophily within communities in the villages

- Future research would involve actually maximizing modularity to break the villages into community networks and testing for homophily within those communities

Acknowledgements

- Dr. Jukka-Pekka Onnela
- Patrick Staples
- MIT Economics Department
- Heather Mattie
- Tonia Smith
- Dr. Rebecca Betensky

Sources

Newman, M. E. (2010). *Networks: an introduction*. Oxford: Oxford University Press.

Newman, M. E. Modularity and community structure in networks.