Grouping Badger Social Networks

Karl Menzel
Biology Department, Reed College
Portland, Oregon
menzelk.edu

ABSTRACT

This ACM-style template describes how to typeset pseudocode as well as write common mathematical symbols. **Copy this project and start by modifying the title, author, etc.** There are also very useful URLS on Moodle for more information.

Keywords

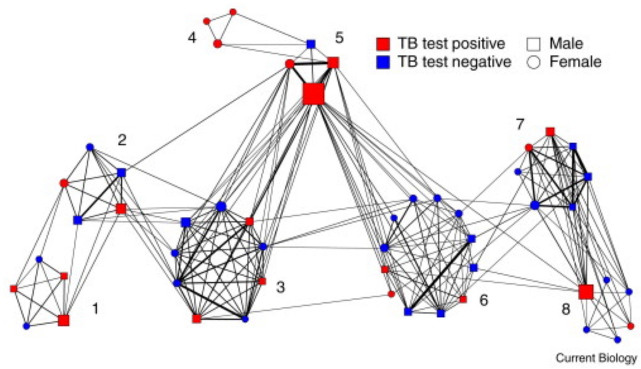
pseudocode, algorithms, math, LaTeX

1. MODIVATION

Animals of some species tend to not be randomly distributed throughout space, instead animals tend to be group up into populations and sub populational groups including social groups. Understanding these social groups can be important for understanding things such as the distribution of disease within the population [4], the social interactions and relationships [**?**], and interpopulation interactions [**?**]. Computational methods using graphs methods have proven to be useful when identifying and social groups and subgroups in highly dynamic free ranging systems [3, 2]. These clustering studies have used weighted graphs created by some measure of association whether by observational studies or by radio collars. Weber et. al. (2013) does not implement clustering on their badger social network, but they use measures of flow-betweennes which are used on some clustering methods.

Most betweennes meausures just use some sort of shortest path method when calculating betweenness. This can leave out some important information about other interactions where information can flow. Information can flow along longer or more circuitous paths and so this should be accounted for in out measure of betweennes [1]. Flow-betweenness does account for this, keeping track of all of the flow that can get between node when calculating betweenanes.

Understanding the social groups of these badgers as well as understanding how flow-betweennes, within and between groups, plays into their social groups.



t

Figure 1: Figure form weber et. al. (2013) ilistrating the social interactions of badgers and distrobution of TB

2. METHODS

2.1 data

The data is an updated version of the data used in Weber et. al. (2013). This data was requested by Anna Ritz in the early fall of 2016.The data is held in two files BadgerInfo.txt and BadgerMatrix.txt.

- **BadgerInfo.txt** This contains demographic information about the badgers in in the study. It contains four columns: badger name, sex, whether or not they are infected with tuberculosis, and the social group the badger is a part of.
- **BadgerMatrix.txt** This contains the interaction data of the badgers. This matrix is constructed as usual with each cell being the interaction between the row and column badger. The number in the cell denotes the amount of time that the two badgers were in close contact

2.2 Pseudocode

The goal is to find the flow-betweenness centrality of all of the nodes in the graph. To do this I use alorgithm 2 to get the normalized amount of flow calculated though each pair of points. This algrorithm uses the Fodr-Fulkerson method to caluclate individual flows though nodes. I seperate the flows into four different sections which use different subsets

Bio331 Fall 2016, Reed College, Portland, OR

ACM ISBN X-XXXXX-XX-X/XX/XX.
DOI: XX.XXX/XXX_X

depending on which social group the node in question and to the sorce and target node are from. First, c_{total} inlcudes all of the possible pairing of source and sink nodes. Seccond c_{btwn} consists of sorce and target are not from the same socal group. The next two are if the source and the target are in the same social group: c_{inter} has the node in question in that same social group, c_{out} has the node in question in a different social group. Algorithm 2 interates between all unique combinations of sources and targes with $(u, v) = (v, u)$ and at each combination finds the normalized flow for all the nodes using 1. It then addes those flows to the centrality group that they are a part of.

The Ford-Fulkerson method (Algorithm 1) works by simulating putting flow throught the graph. It does this by keeping track of a residual network (G_f) which is a representation of how much more flow can go through each of the edges. The algorithm goes until there is not path from the source to the sink where all of the edges > 0 in the residual graph. In other words this is untill there can be no more flow to the target from the source. Each iteration the algorithm find a path using depth first search on G_f, then it adds the minimun edgec_f value to each of the flows and the recalculates G_f. This is like sending the most possible flow though that path which is restricted by edge with the lowest flow. Once there are no more paths in G_f from the source to the target it will return the flows for each of the edges in G.

I also used the Girvan-Newman method to cluster the nodes based on both the edge flow betweenness and the raw edges weights. The Girvan-Newman method works by finding the betweenness of each of the edges. It does this by finding the shortest path between every pair of nodes and then find the number of these paths that each edge is in, then it removes the edge with the highest betweenness. Then the betweennes of all the effected edges are recalculated and another edge is taken off. To find a certain number of groups, those steps are repeated until there are the desired number of connected components.

3. RESULTS

This data consists of all of the interactions of 52 badgers. This population of badgers consists of 9 social groups which are defined by the setts (tunnel system) where they live. Following are statistics that represent the proportion of these interaction that are between members of the same social group and between members that are not in the same social group as well as aggregates for each social group. In table 3 Number of Edges on Nodes in Group is found by $|(i, j) : (i, j) \in E$ and $i \in S_i|$ given nodes i and j in graph G = (V,E) and social group s_i and the next column is the proportion of those edges that had one edge outside the social group.

Figure 2 visualizes the full badger network with my calculated variables. The node sizes are a scaled version of node between group flow betweeness. The edge thicknesses represent the edge flow-betweenness. The badger network was clustered by the Girvan-Newman method using both the raw time data and the edge flow betweenness as edges weights. Figure 4 visualizes the clustering for the raw time edge weights clustered into eight different social groups. Figure 3 visualizes the clustering for flow-betweennness edge weights clustered into eight social groups.

Algorithm 1 Ford-Fulkerson method

Inputs: A network $G = (V, E)$ with flow capasity c, source s, and target t
Outputs: Flows $f(u, v)$ for all $(u, v) \in E$ between s and t
$f(u, v) \leftarrow 0$ for all edges (u, v)
while there exits a path p in G_f : $c_f(u, v) > 0$ for all edges $(u, v) \in p$ **do**
  find $c_f(p) = \min(c_f : (u, v) \in P)$
  for each edge $(u, v) \in p$ **do**
    $f(u, v) \leftarrow f(u, v) + c_f(p)$
    $f(v, u) \leftarrow f(v, u) - c_f(p)$
  end for
end while
return f

Algorithm 2 Flow-Betweenness Centrality

Inputs: $G(E, V)$ and edge weights w social groups s
Outputs: flow-betweenness centrality for nodes
$c_{inter}(n) \leftarrow 0$ for $n \in E$
$c_{btwn}(n) \leftarrow 0$ for $n \in E$
$c_{out}(n) \leftarrow 0$ for $n \in E$
for $k \in V$ **do**
  for $j \in V : (j, k) \in E$ and $j < k$ **do**
    $f(u, v) =$ Ford-Fulkerson(G, w, j, k)
    for n $\in V : n \neq j$ and $n \neq k$ **do**
      $c(n) = \sum_{o \in N(n)} \frac{|f(n,o)|}{2(\sum_{o \in N(k)} |f(k,o)|)}$ where $N(n)$ are n's neigbors)
      if $s(j) = s(k)$ **and** $s(j) = s(n)$ **then**
        $c_{inter}(n) = c_{inter}(n) + c(n)$
      else if $s(j) = s(k)$ **then**
        $c_{out}(n) = c_{out}(n) + c(n)$
      else
        $c_{btwn}(n) = c_{btwn}(n) + c(n)$
      end if
      $c_{total}(n) = c_{total}(n) + c(n)$
    end for
  end for
end for
return $c_{total}, c_{inter}, c_{btwn}, c_{out}$

4. DISCUSSION

The clustering was able to find some of the groups is some of the attempts of clustering, but overall it was not able to cluster the badgers into social groups based on either

Table 1: Summary of whether edges that start in a social group end in the same social group or another social group.

Group	Num Group Edges	Internal Edges (%)
1	109	66
3	44	55
2	9	67
5	60	20
4	41	49
7	76	53
6	27	74
8	128	72

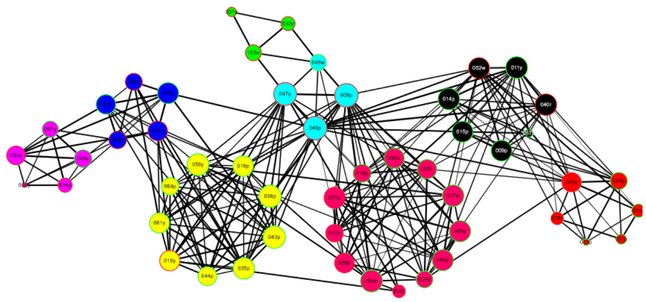


Figure 2: Graph of the interactions of the badgers. Node color represents the badgers social group defined by sett, node size represents between group flow-betweenness, edge width represents edge betweenness.

the raw time edge weights or by the edge flow-betweenness weights. The clustering was not much better when there was more or fewer clusters. Instead of breaking up the sett based groups it merely just split off different chunks of other groups into their own small groups. The between group node flow betweenness alone did not seem to be a great predictor of whether or not it would end up in a group with its other sett mates. Some indiviauls such as the small black node in figure 4 and the small orange node in figure 3 did not manage to stay with their groups. Most of the nodes in the yellow and orange groups have high between group flow-betweenness but were clustered with their group.

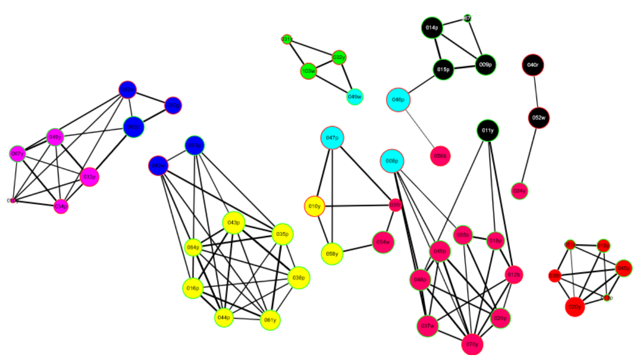


Figure 3: Graph of the interactions of the badgers. Node color represents the badgers social group defined by sett, node size represents between group flow-betweenness, edge width represents edge betweenness. Edges were removed vie the Girvan-Newman method to create 8 groups to create 8 groups seen as the 8 connected components seen here.

The clustering was able to cluster some groups pretty well most of the time. The yellow, green, and orange group both had a large number of their members in one group. Some groups did not cluster well at all such as the cyan group and the black group. There seems to be some sort of correlation between this and the percent of internal nodes (Table 3). The yellow and the green groups both had high percent of internal nodes and the cyan and the black group had lower percentages of internal nodes.

Ultimately, I think there are some things missing in the

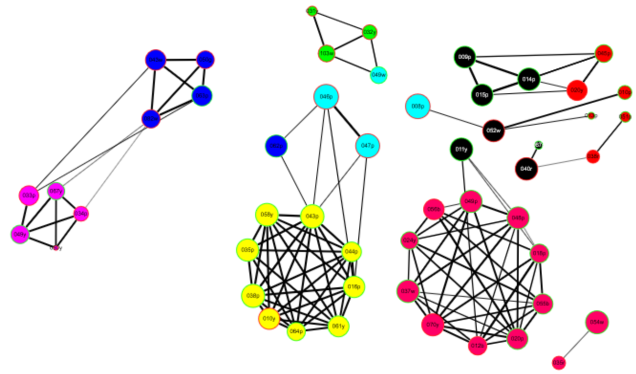


Figure 4: Graph of the interactions of the badgers. Node color represents the badgers social group defined by sett, node size represents between group flow-betweenness, edge width represents time in contact with. Edges were removed vie the Girvan-Newman method to create 8 groups to create 8 groups seen as the 8 connected components seen here.

time of interaction data that is needed to cluster the badgers into the correct setts. Webber et. al. (2013) found that badgers with tuberculosis interacted less with their own sett and more with other setts depending on the season. This could cause badgers to groups into the wrong cluster. The setts are also spaced differently geographically, this would make it harder or easier for different badgers to interact with any of the setts. Incorporating this kind of information, possible when calculating edge weights, may make the clustering better for these badgers but it may also make the clustering too specific to this system and not general enough to be useful anywhere else.

5. REFERENCES

[1] Linton C Freeman, Stephen P Borgatti, and Douglas R White. Centrality in valued graphs: A measure of betweenness based on network flow. *Social networks*, 13(2):141–154, 1991.

[2] Maija K Marsh, Steven R McLeod, Michael R Hutchings, and Piran CL White. Use of proximity loggers and network analysis to quantify social interactions in free-ranging wild rabbit populations. *Wildlife Research*, 38(1):1–12, 2011.

[3] Gabriel Ramos-Fernández, Denis Boyer, Filippo Aureli, and Laura G. Vick. Association networks in spider monkeys (ateles geoffroyi). *Behavioral Ecology and Sociobiology*, 63(7):999–1013, 2009.

[4] Nicola Weber, Stephen P Carter, Sasha RX Dall, Richard J Delahay, Jennifer L McDonald, Stuart Bearhop, and Robbie A McDonald. Badger social networks correlate with tuberculosis infection. *Current Biology*, 23(20):R915–R916, 2013.