Project Title:

EdgeLap: Identifying and discovering features from overlapping sets in networks

Names and Email Addresses:

Jessica Wong (jhmwong@cs.ubc.ca)
Aria Hahn (hahnaria@gmail.com)
Sarah Perez (karatezeus21@gmail.com)

Description of Domain, Data, and Task:

Microbes are single celled organisms that can be found everywhere in the world from the air to the soil on the ground. Generally, microbes will be found in extremely close proximity to other species of microbes due to mutual beneficial relationships where one species produces something that another species requires for survival; these groups of microbes are called a community. The concept of microbial species is not yet well defined thus instead microbes are grouped into operational taxonomic units (OTUs) based on their DNA. In the rest of this proposal, microbial species will be referred to as an OTUs. Each OTU can be identified based on five different taxonomic categories: kingdom, phylum, class, order, and genus. Taken together, these five categories make up the taxonomy of the microbe.

A widespread question in microbiology is trying to identify interactions between OTUs, kingdoms, phylums, class, orders, and genus in multiple environments and communities. For this project, I am working with over 700 DNA samples of microbial communities that were harvested from soil samples obtained from three sites at six different geographical locations across Canada and the United States. At each site, microbial DNA was extracted from soil in four forest plots that underwent different forest harvesting treatments four different ways (unmanaged, mild, moderate, and heavy). The extracted DNA was then sequenced and used to to identify the OTUs present within each sample. In order to identify potential microbial interactions, co-occurrence networks were constructed for: all samples, each geographic location, each site within each geographic location, and each forest treatment within each geographic location resulting in a total of 52 networks. The network visualization and calculation of network properties has been completed outside of the scope of this project.

Each soil microbial community network is represented in the form of two files. The first file lists what OTUs are present in a sample (these are the data items or nodes in a network). Each node will also have five attributes (kingdom, phylum, class, order, and genus) that will describe its taxonomy. The second file lists which OTU is correlated with another OTU (there are the links between nodes in a network); these correlations can be positive or negative. My dataset consists of approximately fifty soil microbial networks constructed from over 700 sample. On average, there are about three thousand correlations in each network. As I am not an expert in this domain, I will be asking Aria Hahn and Sarah Perez about any further questions regarding the underlying biology that I may encounter during the course of this project.

The basic task I want to accomplish is to identify and locate common links across any combination of networks in my dataset. I am not focusing on just showing the links that appear in all of the networks; I will also be trying to visualize the sets of links that are common to a subset of the networks I am looking at. The goal is to generate a hypothesis about the types, number and taxonomy of correlations found between OTUs in different communities.

Personal Expertise:

I have some microbiology knowledge as I did my undergrad in microbiology and computer science. I have also done a microbial diversity lab that focused on these problems of trying to find correlations between interactions OTUs in soil and sewage so the general idea of this project is not a foreign concept to me. I really enjoyed the lab when I took it so when Aria pitched her project, it sounded like an interesting problem I wanted to help tackle.

Proposed Vis Solution:

As mentioned in the “Description of Domain, Data, and Task” section, I want the visualization to identify and describe common links across any combination of networks in my dataset. I am trying to build a tool that will take in two to seven networks and will allow the user to identify all sets of overlapping edges in any combination of networks (i.e., if I am examining three networks, I want to see two sets of edges: one set will be the edges that appear in all three networks, while the other set will show edges that only appear in any two of the three networks). The visualization I will use for this task will be a combination of radial sets [1] as well as multiple windows.

<table>
<tr><th>Idiom</th><th>EdgeLap</th></tr>
<tr><td>WHAT – data</td><td>Several networks where nodes are OTUs (microbial species) and links are positive/negative correlations. The networks are undirected, and asymmetric. There is no guarantee on the size of the networks in relation to each other nor does the shared set of links have to be non-empty.</td></tr>
<tr><td>WHAT – derive</td><td>The links common to any user specified combination of networks given network attributes.</td></tr>
<tr><td>HOW– encode</td><td>
<table>
<tr><th>Data Attribute</th><th>Mark or Channel</th></tr>
<tr><td>Network</td><td>Glyph</td></tr>
<tr><td>Quantity/distribution of links in a network</td><td>● Length (histogram inside network glyph describes quantity/distribution of links using length)
● Position on a common scale</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td></td><td>
<table>
<tr><td></td><td>● Colour</td></tr>
<tr><td>Links in overlap</td><td>● Line mark
● Connection mark</td></tr>
<tr><td>Link Attributes</td><td>● Saturation
● Glyph
● Colour</td></tr>
</table>
</td></tr>
<tr><td>WHY – tasks</td><td>
<table>
<tr><th>Actions</th><th>Targets</th></tr>
<tr><td>Discover</td><td>Features and similarities/differences between networks</td></tr>
<tr><td>Explore</td><td>Distribution of links</td></tr>
<tr><td>Identify</td><td>● Similarities between networks
● Features in the grouping of links</td></tr>
<tr><td>Compare</td><td>Networks</td></tr>
</table>
</td></tr>
<tr><td>Scale capacity</td><td>● Working with 2-7 networks at any one time
● Each network has hundreds of nodes (100-300)
● Each network has, on average, 3000 links, but can range from ~100 to ~6000</td></tr>
</table>

Table 1: A breakdown of the task and data abstraction for EdgeLap based on principles discussed in [5].

The main window of the visualization (the one on the left) will be responsible for displaying radial sets which will depict the overlapping sets of edges between all networks currently being investigated. For each network depiction in the radial set, there will be a set of aligned histograms that are meant to give the user a general sense of how many items in the network participate in various network relationships. The outermost bar in the network depiction (the longest one) will represent the total number of links in all the networks involved in the current comparison (i.e., the sum of all links from each network involved in the radial set). The next bar inward will show the total number of links specific to the given network; this bar will be subdivided into two portions of different saturation with the darker portion indicating how many links in the network are not found in other networks and the lighter portion indicating how many links in the network are found in other networks. The next bar inward will show the number of edges in the network that are shared with exactly one other network. In other

words, if OTU1 and OTU2 had a correlation that appeared in both network 1, 2, and 3, it would not be counted in this histogram bar; if an edge between OTU1 and OTU2 existed in network 1 and 2 but not 3, then OTU1 and OTU2 would be counted in the histogram bar. The next bar inward will show the number of edges in the network that are found in exactly two other networks. As the histogram bar gets closer to the centre of the radial set, it will denote how many edges are shared with an increasing number of networks. The histogram alignment is meant to help the user quickly discern general properties about the number of links that shared with other networks.

For m networks, there are 2^m relationships each network can participate in. Since lines denote relationships between networks, the natural consequence is that the number of lines present inside the radial set is very large and hence, hard to decipher. I decided to use a colour saturation to help users differentiate between lines that depicted edges that were shared between different numbers of networks. Since the users are more interested in cases where more networks share a common edge, lines that denote these relationships use more saturated colours. Likewise, sets that depict edges shared between a smaller number of networks are represented by lines that use less saturated colours (see figure 2 and 3). Thickness was not considered as an encoding due to the sheer number of lines that could result from the radial set; thickness would quickly become indiscernible given enough lines.

In order to help users differentiate between which sets are participating in multi-network relationships (any relationships involving more than two networks), lines that denote an overlapping set of edges between more than two networks are connected to a circle. The size of the circle will depend on the cardinality of the overlapping set of edges; the larger the cardinality, the bigger the circle will be. This encoding was chosen to help users pick out which networks were involved in a specific multi-network relationship as well as to help users gain a preliminary intuition of which networks might share many interactions without having to create another radial set to investigate further. The circle encoding can be one of six possibilities:

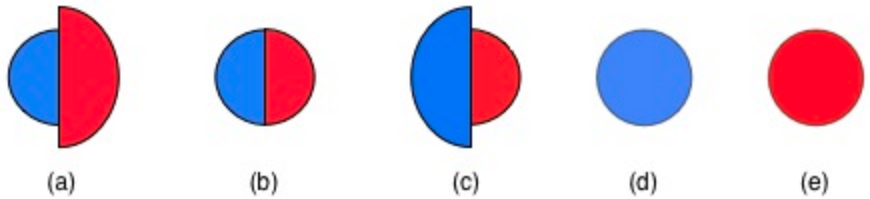

Figure 1: The circle glyphs will represent the relationship between the cardinality of the set of positive correlations and the cardinality of the set of negative correlations. (a) There are less positive correlations than negative correlations, (b) there are an equal number of positive and negative correlations, (c) there are more positive correlations than negative correlations, (d) there are only positive correlations, and (e) there are only negative correlations.

Ideally, there would be some way to filter the number of edges shown in the radial set to reduce visual clutter and for the user's ease in exploring the different overlapping sets. A

possible method would be to filter the radial set edges by the number of networks participating in each edge. For example, I could filter the full set of results to only the radial set edges from any three networks rather. However, due to time constraints, this is will not a feature that will show up in the visualization.

The view on the right of the radial set will hold the list of networks—this would list all the soil sample files that the visualization application can find. This is simply an interface for the user to pick and choose which networks he/she wants to examine.

The checkboxes located underneath the list of networks are for filtering the results of the overlapping sets. In figure 1, the results could hold OTU interactions that exist but do not match in terms of being a positive or negative correlation. For example, OTU1 and OTU2 could have a positive correlation in network 1 but a negative correlation in network 2. When I examine network 1 and 2 together, I will see OTU1 and OTU2 as a common interaction that occurs in both networks. While knowing that the interaction is common among networks is useful, it would also help users if there was some way to filter those results out so that only interactions which are of the same “type” are displayed. This is what the “Unidirectional correlations” checkbox is for. Likewise, the “Positive Correlations” and “Negative Correlations” checkboxes filter the results of the overlapping sets. In the case of “Unidirectional correlations”, as both positive and negative correlations will be represented in the radial set with the circle glyph shown in figure 1. If time permits, there will also be extra functionality that would allow users to examine the networks for correlations based on the taxonomy of the OTUs that are involved in interactions.

Users should also be able to see a classical network view of the overlapping set in question (see figure 5). Each of the individual networks are often too large to be put into a classical network view but the subset of the networks is generally small enough that a classical network view would be helpful as it could clearly indicate which OTUs are involved in correlations. It would also indicate the nature of those correlations as being positive (blue) or negative (red). Note that this colour scheme is identical to the one used for the circle glyph. This window does not need to be in a multiple view relationship with the radial set because according to the users, they would not really need to refer to the radial set when examining the overlapping set. In order to maximize the use of pixels during each stage of the users’ workflow, the classical network view of the overlapping set will appear in another window.

There will also be an export function for the user to export files from the visualization: 1) the a network view of what OTUs are involved in the set of overlapping edges (see figure 6), 2) the radial set as it is currently displayed on the visualization, and 3) a list of all the overlapping sets that occur in that radial set.

Usage Scenario:

To Begin:

To begin, the user would select the files that he/she wanted to examine in the radial set. After making the selections, the user would have to click the Add button located on the bottom of the list of files to have his/her file selections created into a radial set. Once the Add button has been clicked, the user can start examining the radial set generated on the left.

Examining the Radial Set:

The main window (the one of the left) will display the radial set based on the networks the user has specified. If the user wishes to examine a specific overlapping set in the radial set further, he/she can click on the circle glyph representing that overlapping set to have information about that set appear in another window (figure 5).

Examining the Network Attributes:

The users have also expressed an interest in knowing general numbers about the networks involved in the radial set (e.g., how many links does the radial set have in total, how many of those links are from a specific network, how many links in a network participate in a relationship with other networks etc.). A general sense of this information can be obtained by looking inside each network mark at the aligned histograms. The green histogram represents all the links in the network and the other histogram bars will represent information about the links specific to the network in question. If the user wishes to find exact numbers or to examine the histograms more carefully, then he/she can click on the network mark (figure 6).

Exporting:

After investigation, if the user wishes to export the radial set image along with the common edges found in that radial set, he/she can click Export. If none of the overlapping sets has been selected when Export is clicked, then it is assumed that the user only wants to find the overlapping set of edges that appear in all the radial set networks. However, if the user has selected a certain overlapping set when clicking Export, then the information pertinent to that selected set will be written to the output files. In either case, there will be three files outputted: 1) a text file that lists all the common edges found in the network, 2) a SVG of the classical network view of the overlapping set, and 3) a SVG of the radial set.

Illustrations of the interface:

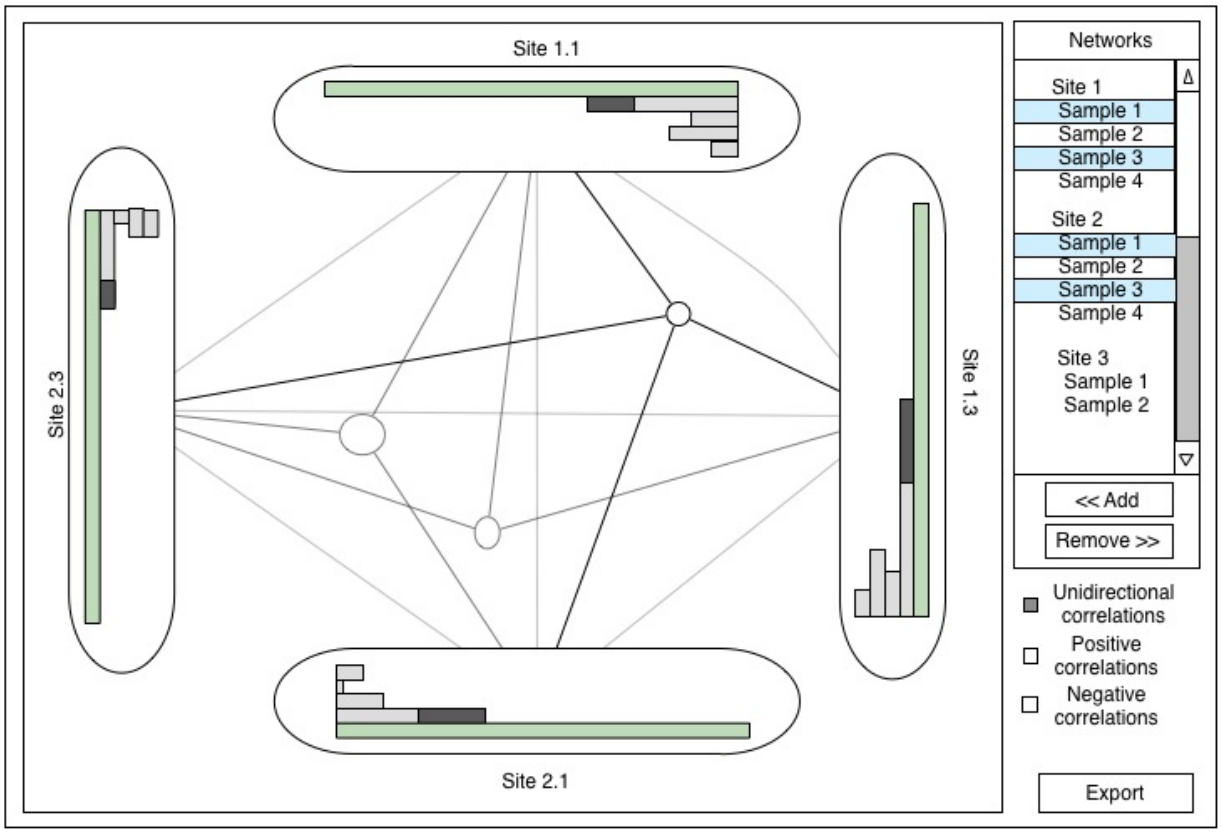

Figure 2: What the visualization would look like if we were examining all correlations between the networks regardless of whether the OTU correlations that occur in the sets are all of the same “type”.

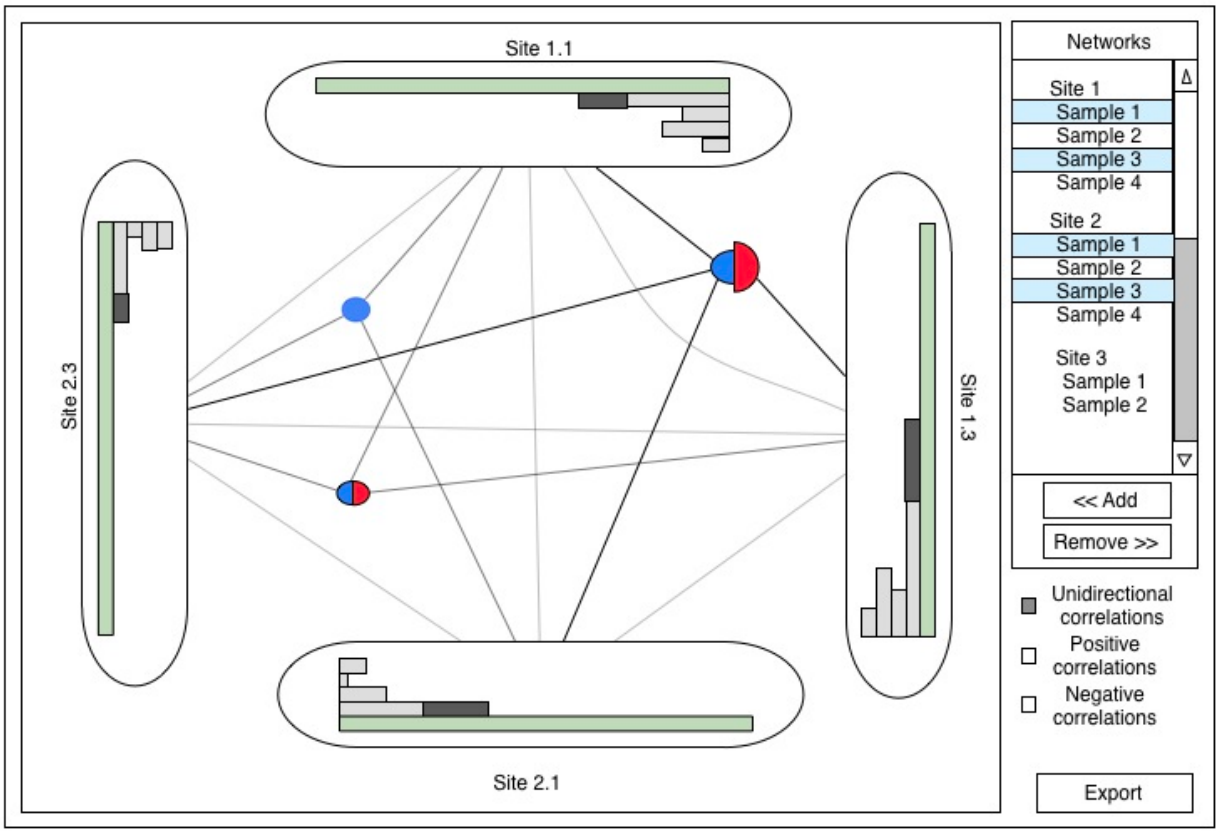

Figure 3: What the visualization would look like if we imposed constraints on the set of OTU correlations that occur in the sets to be all of the same ”type”. The circles in the middle of the radial set represent the set of overlapping edges between the networks that connect to them. The set of positive correlations within the set of overlapping edges are represented by the blue circle and the set of negative correlations are represented by the red. The size of the circle as a whole indicates how large the overlapping edges are while the size proportion between the red and blue half circle indicates whether the set of positive correlations are lesser than, equal to, or greater than the set of negative correlations.

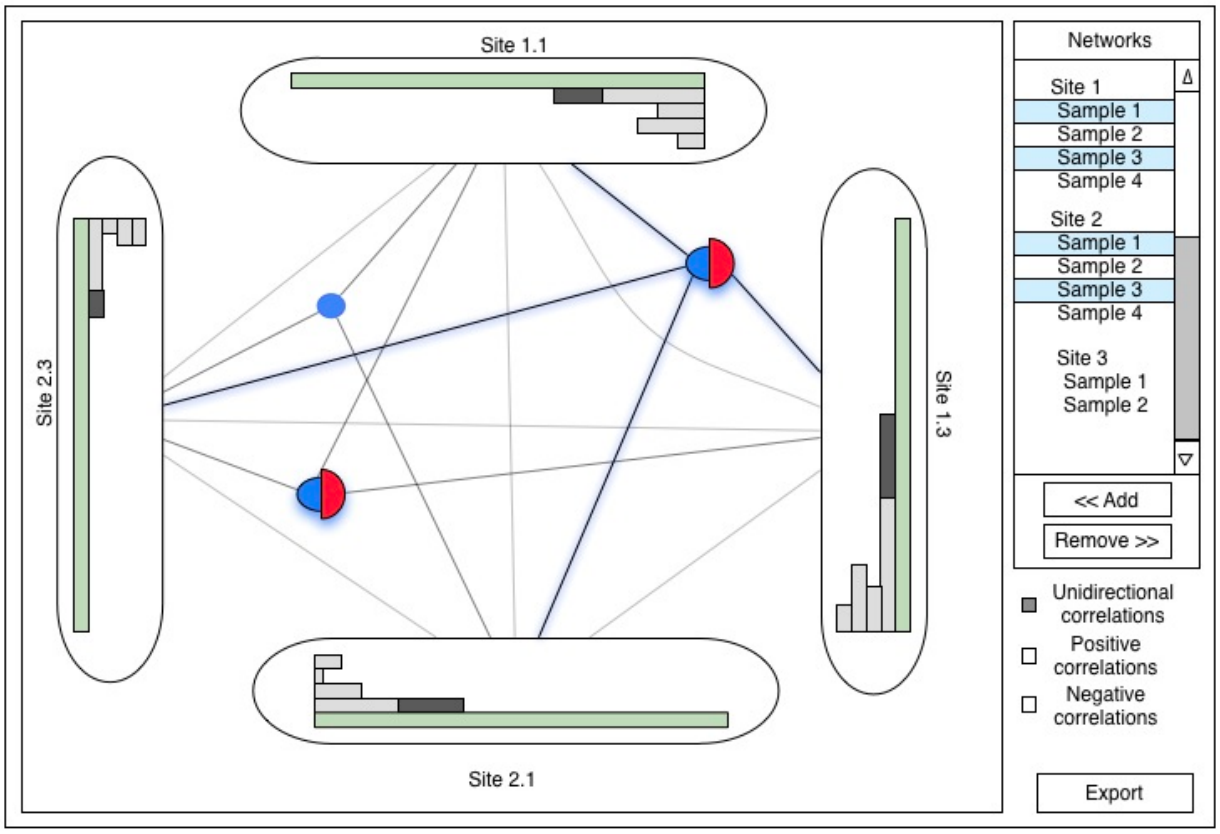

Figure 4: The user has selected a specific overlapping set to examine. If the user decides to double click the circle, the set of overlapping edges from the networks connected to the circle will be displayed in a classical network view (see figure 5) in a new window.

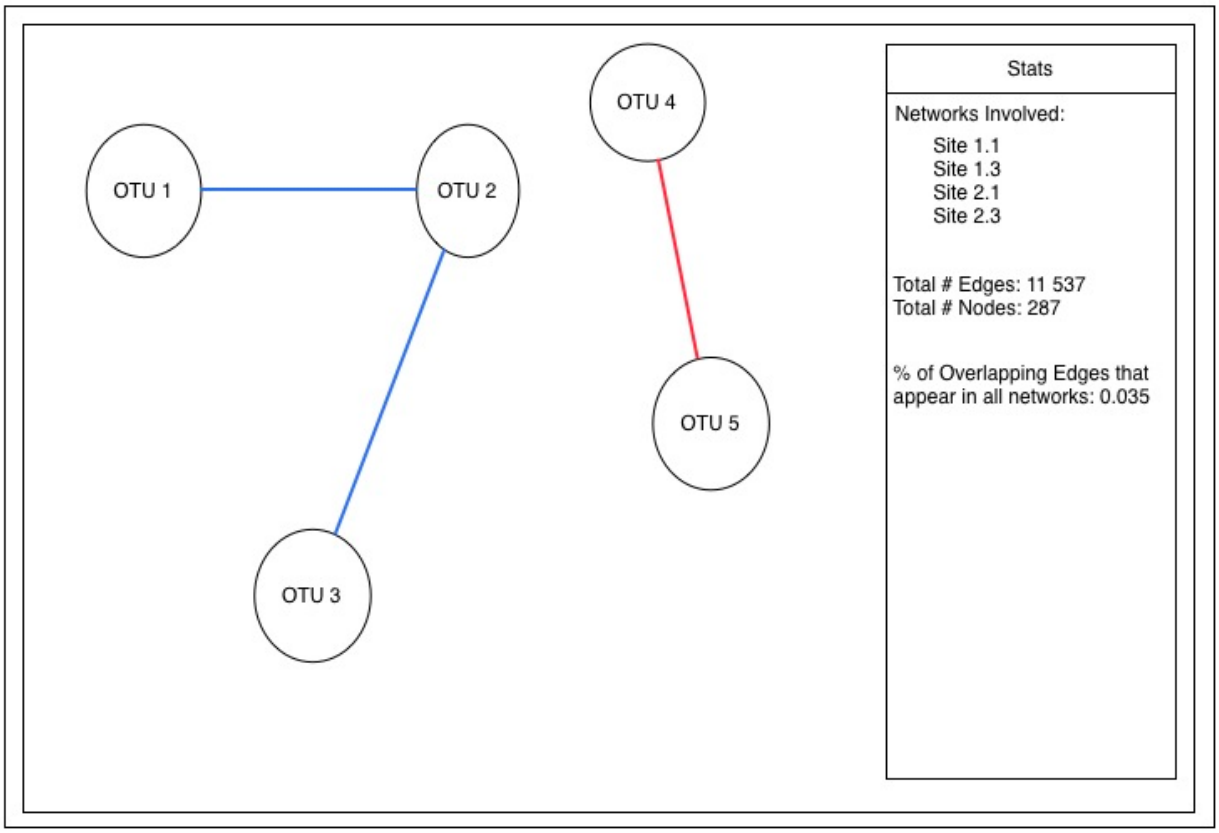

Figure 5: The classical network view of the overlapping edges found from the Radial Set if the user had double clicked the selected circle glyph in figure 3. The numbers at the side of the network list out basic numbers related to the network and edges involved in overlapping sets.

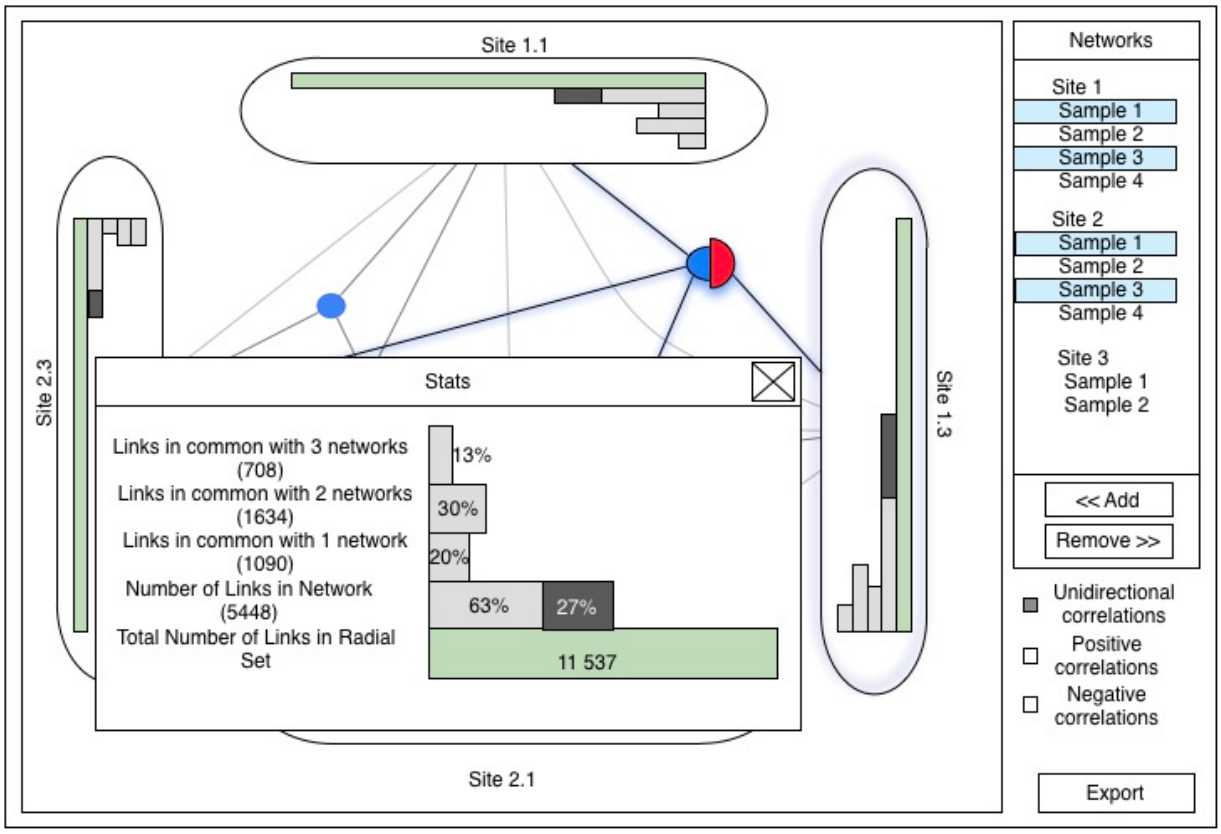

Figure 6: If the user decides to double click on one of the sets in the radial set, he/she can view how many links the set has in common with other networks as well as how many links the set contributes to the total number of links.

Proposed Implementation Approach:

- Processing/Javascript to create the visualization itself
- Debating between Java to parse files into a form that would speed up the visualization
    - C would be faster but I am rusty on my C and the time required to code that up may be significantly longer than if I did it in Java. Due to the amount of work implementing radial sets is expected to take, I may just write the data parser in Java.
    - Also will do some quick research to see if Javascript is good at doing tasks like this.

Milestones and Schedule:

Week 1 (Nov. 1 – Nov. 7)

- Get familiar with the tools/languages discussed in the “Proposed Implementation Approach” (Nov. 1 – Nov. 6)
- Get the code to parse the data up and running (Nov. 7)

Week 2 (Nov. 8 – Nov. 14)

- Get the main GUI with different windows shown (Nov. 12)
- Get the list of files to load into the side window (Nov. 14)
- Get the select/deselect network function to work (Nov. 14)

Week 3 (Nov. 15 – Nov. 21)

- Work on the display of the radial set

Week 4 (Nov. 22 – Nov. 28)

- Finish display of radial set (Nov. 26)
- Get the line selection and display the classical network view of the overlapping set (Nov. 28)

Week 5 (Nov. 29 – Dec. 5)

- Populate the stats window for when a user double clicks a set (Dec. 2)
- Getting Export to work (Dec. 3)
- If I am done the original scoped project by Dec. 3, then I will consider adding in functionality where we can look for shared links based on taxonomy
- Debugging
- Schedule catch-up

Week 6 (Dec. 6 – Dec. 12)

- Debugging (Dec. 11)
- Working on presentation (Dec. 11)
- Start working on write up

Week 7 (Dec. 13 – Dec. 15)

- Finish write up (Dec. 15)

Previous Work:

This visualization is heavily inspired by the radial sets discussed in [1]. I am currently in the process of trying to find other applications that have used radial sets; hopefully, there will be some documentation about challenges they have faced so I can be more prepared going into the implementation phase of the project.

I will also be looking into current commercial tools [2, 3, 4] that are being used to examine individual biological networks to see if there are any good visual encodings or interactional idioms that can be adapted into this visualization.

References:

[1] B. Alsallakh, W. Aigner, S. Miksch, and H. Hauser. (2013). Radial Sets: Interactive Visual Analysis of Large Overlapping Sets. *IEEE Transactions on Visualization and Computer Graphics (Proc InfoVis 2013) 19(12),* 2496-2505.

[2] Hive Plotter. (2011). http://www.hiveplot.net/

[3] Cytoscape. (2001). http://www.cytoscape.org/what_is_cytoscape.html

[4] Gephi. (2008). http://gephi.github.io/

[5] T. Munzner. Visualization Analysis and Design. (2014). A K Peters Visualization Series. CRC Press.