3.1 Sci2 Algorithms and Tools

Loading Data (Files)

- ***Facebook***
  - Access Token - Allows the user to authenticate with Facebook, simplifying the process of using the other Facebook algorithms.
  - Facebook Friends Data - Pulls the name, user ID, latest update, gender, current location, hometown, birth date, interests, religion, political views, relationship status, and attended events of all the friends in a user's network.
  - Mutual Friends - Returns a CSV with every single connection between a user and his or her friends, and calculates the number of mutual friends for each connection.
- ***Google Scholar***
  - Google Citation User ID Search Algorithm - The Google citation user ID is a web reader algorithm that retrieves the Google citation for the specified author.
  - Attach Citation Indices from Google Scholar - Reads the citation indices from the Google scholar profile for a given set of authors and creates a csv file with the corresponding citation indices.
  - Attach Citation Table from Google Scholar - Reads citations in the Google scholar profile for a given author and returns the citation information in the form of a series of tables for all the queried Google Citation user ID
- ***Flickr Reader***
  - Flickr reader - Reads a list of Flickr User IDs from a CSV file loaded into Sci2 and attempts to gather the image URLs of every image those users uploaded
- ***T*witter Reader**
  - Twitter reader - Reads a list of Twitter handles from a CSV file and attempts to pull their recent status updates and information about the individual status updates. The user may also specify hashtags.

Data Preparation

- ***Text Files***
  - Convert to Generic Publication - A plugin which takes as input a path to a .hmap properties file, then uses the header mapping in the file to standardize the headings in a table.
  - Remove ISI Duplicate Records – Removes duplicate publications form ISI records based on ISI Unique ID attribute.
  - Remove Rows with Multitudinous Fields – Removes rows having at least N entries within a given field.

  - Extract Directed Network – Creates a directed network by placing a directed edge between the values in a given column to the values of a different column.

- Extract Bipartite Network – Creates an unweighted bipartite network by placing a directed edge between the values in a given column to the values of a different column.
- Extract Paper Citation Network – Extracts an unweighted directed network from papers to their citations.
- Extract Author Paper Network – Extracts an unweighted directed network from authors to their papers.

- Extract Co-Occurrence Network – Extracts a network from a delimited table.
- Extract Word Co-Occurrence Network – Creates a weighted network where each node is a word and edges connect words to each other. The strength of an edge represents how often two words occur in the same body of text together.
- Extract Co-Author Network – Extracts a weighted network with authors as nodes and edge weights as the number of times those authors co-wrote a paper.
- Extract Reference Co-Occurrence (Bibliographic Coupling) Network – Extracts a weighted network from a Paper Citation network, with papers as nodes and edge weights as the number of citations two papers share.

- Extract Document Co-Citation Network – Extracts a weighted network from a Paper Citation network, with papers as nodes and edge weights as the number of times two papers are cited together.

- Detect Duplicate Nodes – Cleans graph data by detecting and preparing to merge nodes that are likely to represent the same entity.
- Update Network by Merging Nodes – Creates a new network by running the algorithm with both the Merge Table from "*Detect Duplicate Nodes*" and the original network selected.

Preprocessing

- ***General***
  - Extract Top N% Records – Returns the top N% rows of a table by selecting the percentage of rows to keep and column to sort by.
  - Extract Top N Records – Returns the top N rows of a table by selecting the number of rows to keep and column to sort by.
  - Aggregate Data – Summarizes the input table by column, allowing the aggregation of values such as "Cited Reference Count," "Number of Pages," "Publication Year," "Times Cited," as well as values represented by many other delimiters.
- ***Temporal***
  - Slice Table by Time – Slices a table into groups of rows by time.
- ***Geospatial***
  - Extract ZIP Code – Extracts a ZIP code from a given address.
- ***Topical***
  - Reconciled Journal Names - This algorithm maps the given journal names to the equivalent journal names in the UCSD Map of Science standard.
  - Lowercase, Tokenize, Stem, and Stopword Text – Replaces spaces and punctuation from a field with a standard delimiter of the user's choosing.
- ***Networks***
  - Extract Top Nodes – Extracts the top N nodes from a graph, based on a given attribute.
  - Extract Nodes Above or Below Value – Extracts nodes with an attribute above or below a certain value.
  - Delete Isolates – Removes nodes which are not connected to any other in the graph.
  - Extract Top Edges – Extracts the top N edges from a graph, based on a given attribute.
  - Extract Edges Above or Below Value – Extracts all edges with an attribute above or below a certain number from a graph.
  - Remove Self Loops – Removes edges whose source and target nodes are equivalent from a graph.
  - Trim by Degree -- Deletes edges at random until each node has at most N edges.
  - MST-Pathfinder Network Scaling – Prunes a network using the MST-Pathfinder algorithm.
  - Fast Pathfinder Network Scaling – Prunes a network using the Fast Pathfinder algorithm.

  - Snowball Sampling (N nodes) – Picks a random node and traverses its edges iteratively until N nodes are extracted.
  - Node Sampling – Extracts N random nodes and their intervening edges, and then deletes isolates.
  - Edge Sampling – Extracts N random edges and their target and source nodes.

  - Symmetrize-- Turns a directed network into an undirected network.
  - Dichotomize – Trims edges above, equal to, or below a certain value.
  - Multipartite Joining – Joins a multipartite graph for one node type across another node type.

  - Merge 2 Networks – Merges identical networks. Emphasis is put on edge attributes merging.

Analysis

- ***Temporal***
  - Burst Detection – Determines periods of increased activity in a table with dates/timestamps.
- ***Geospatial***
  - Geocoder – Converts place names to latitudes and longitudes.
  - Congressional District Geocoder - Converts the given **9-digits U.S. ZIP codes (ZIP+4 codes)** into its congressional districts and geographical coordinates (latitude and longitude).
  - Yahoo! Geocoder - Converts place names and addresses into latitudes and longitudes (requires Yahoo! API Key)
- ***Topical***
  - Burst Detection – Determines periods of increased activity in a table with dates/timestamps.
- ***Networks***
  - Network Analysis Toolkit (NAT) – Calculates basic network statistics, such as number of nodes (and isolated nodes), node attributes, number of edges, presence of self loops and parallel edges, average degree ("total," "in," and "out"), strength of component connections, and overall density.
  - ***Unweighted & Undirected***

- Node Degree – Calculates the amount of edges adjacent to a node, and then appends that value to each node.
- Degree Distribution – Builds a histogram of the degree values of all nodes.

- K-Nearest Neighbor – Calculates the correlation between the degree of a node and that of its neighbors, and then appends that value to each node.
- Watts-Strogatz Clustering Coefficient – Calculates the degree to which nodes tend to cluster together, and then appends that value to each node.
- Watts Strogatz Clustering Coefficient over K – Correlates the clustering coefficient and the degree of the nodes of a network.

- Diameter – Calculates the length of the longest shortest path between pairs of nodes in a network.
- Average Shortest Path – Calculates the average length of the shortest path between pairs of nodes in a network.
- Shortest Path Distribution – Builds a histogram of the lengths of shortest paths between pairs of nodes in a network.
- Node Betweenness Centrality – Appends a value to each node which correlates to the number of shortest paths that node resides on. The more shortest paths between node-pairs a certain node resides on, the higher its betweenness centrality.

- Weak Component Clustering – Extracts the N largest weakly connected components of a network.
- Global Connected Components – Calculates the number of connected components or subgraphs with a path between each pair of nodes.

- Blondel Community Detection – Extracts a hierarchical community structure for a large network.
- Louvain Community Detection - Large network community detection
- Louvain Multilevel Refinement Community Detection - Large network community detection
- SLM Community Detection - Large network community detection
- ---
- Extract K-Core – Extracts the kth K-Core from a graph. The kth K-Core is what remains of the graph after every node with fewer than k edges connected to it is removed from the graph recursively.
- Annotate K-Coreness-- Appends to each node the K-Core that node belongs to.

- HITS – Computes authority and hub score for every node.

- ***Weighted & Undirected***
  - Clustering Coefficient – Calculates the degree to which nodes tend to cluster together, and then appends that value to each node.
  - Nearest Neighbor Degree - Determines the average nearest neighbor degree.
  - Strength vs Degree - This algorithm determines the strength distribution.
  - Degree & Strength - Determines the degree and strength of each node.
  - Average Weight vs End-point Degree - Determines the average weight as a function of end-point-degree.
  - Strength Distribution - Determines the strength distribution.
  - Weight Distribution - Determines the weight distribution.
  - Randomize Weights - Redistributes the weights, while keeping the topology invariant.

  - Node Betweenness Centrality - We use Brandes' algorithm to calculate the 'betweenness centrality' for vertices.

  - Blondel Community Detection – Extracts a hierarchical community structure for a large network.
  - Louvain Community Detection - Large network community detection
  - Louvain Multilevel Refinement Community Detection - Large network community detection
  - SLM Community Detection - Large network community detection
    ---
  - HITS – Computes authority and hub score for every node.
- ***Unweighted & Directed***
  - Node Indegree – Appends the number of incoming edges to each node.
  - Node Outdegree – Appends the number of outgoing edges to each node.
  - Indegree Distribution – Builds a histogram of the values of the indegree of all nodes.
  - Outdegree Distribution – Builds a histogram of the values of the outdegree of all nodes.

  - K-Nearest Neighbor – Calculates the correlation between the degree of a node and that of its neighbors, and then appends that value to each node.
  - Single Node In-Out Degree Correlations – Calculates the correlations between indegree and outdegree of a node.

  - Dyad Reciprocity – The ratio of dyads with a reciprocated tie to dyads with any tie.
  - Arc Reciprocity – The ratio of reciprocal edges to total edges.
  - Adjacency Transitivity – The ratio of transitive triads to intransitive triads (triads missing one edge).

  - Weak Component Clustering – Extracts the N largest weakly connected components of a network.
  - Strong Component Clustering – Extracts the N largest strongly connected components of a network.

  - Blondel Community Detection – Extracts a hierarchical community structure for a large network.
  - Louvain Community Detection - Large network community detection
  - Louvain Multilevel Refinement Community Detection - Large network community detection
  - SLM Community Detection - Large network community detection
    ---
  - Extract K-Core – Extracts the kth K-Core from a graph. The kth K-Core is what remains of the graph after every node with fewer than k edges connected to it is removed from the graph recursively.
  - Annotate K-Coreness – Appends to each node the K-Core that node belongs to.

  - HITS – Computes authority and hub score for every node.
  - PageRank – Ranks the importance of a node by how many other important nodes point to it.
- ***Weighted & Directed***
  - Blondel Community Detection – Extracts a hierarchical community structure for a large network.

- Louvain Community Detection - Large network community detection
- Louvain Multilevel Refinement Community Detection - Large network community detection
- SLM Community Detection - Large network community detection
  ---
- HITS – Computes authority and hub score for every node.
- PageRank – Ranks the importance of a node by how many other important nodes point to it, taking into account edge weights.

Modeling

- ### Networks

  - Random Graph – Generates a graph with a fixed number of nodes connected randomly by undirected edges.
  - Watts-Strogatz Small World – Generates a graph whose majority of nodes are not *directly* connected to one another, but are still connected to one another via relatively few edges.
  - Barabási-Albert Scale-Free – Generates a scale-free network by incorporating growth and preferential attachment.

  - TARL (Topics, Aging and Recursive Linking) – Incorporates "aging" to generate bipartite coevolving networks of authors and papers. Can also be applied to other datasets with different aging distribution.

R

- Create an R Instance - This algorithm creates an R instance that is usable from the CIShell environment.
- Run Rgui - This algorithm opens the RGui for an already created R instance.
- Send a Table to R - This algorithm imports a CSV into a running R instance.
- Get a Table from R - This algorithm exports a CSV from a running R instance back onto the Data Manager.

Visualization

- ### General
  - GnuPlot – Used to plot two-dimensional functions and data points in many different formats. For full documentation of this open-source software, please visit http://www.gnuplot.info/documentation.html
- ### Temporal
  - Temporal Bar Graph
  - Horizontal Bar Graph
  - Horizontal Bar Graph (not included version) – Uses csv (tabular) datasets (including NSF grant data and the output of burst detections) to visualize numeric data over time, generating labeled horizontal bars that correspond to records in the original dataset.
- ### Geospatial
  - Proportional Symbol Map – Maps geospatial coordinates as circles that can be size- and color-coded in proportion to associated numeric data.. Result is a PostScript file.
  - Choropleth Map – Color-codes named regions on a geographical map in proportion to associated numeric data. Result is a PostScript file.
  - Geospatial Network Layout with Base Map – Allows for the geospatial visualization of network data, by producing a network file and corresponding blank map.
- ### Topical
  - Map of Science via Journals - The Map of Science is a visual representation of a network of 554 subdisciplines of science (grouped into 13 overarching disciplines) and their relationships to one another.
  - Map of Science via 554 - This visualization works exactly like Map of Science via Journals except instead of taking a collection of journal names as input and mapping them to the 554 fields it directly takes IDs of the 554 fields, which are integers from 1 to 554.
- ### Networks
  - GUESS – Interactive data analysis and visualization tool.
  - Gephi - Interactive data analysis and visualization tool at http://gephi.org.

  - Radial Tree/Graph (prefuse alpha) – A single node is placed at the center and all others are laid around it in a tree structure.
  - Radial Tree/Graph with Annotation (prefuse beta) – A single node is placed at the center and all others are laid around it in a tree structure, with labels.
  - Tree View (prefuse beta) – Visualizes directory hierarchies in a tree structure. Warning: Does not work on Macs.
  - Tree Map (prefuse beta) – Visualizes hierarchies using the Treemap algorithm. Warning: Does not work on Macs.
  - Force Directed with Annotation (prefuse beta) – Sorts randomly placed nodes into a more aesthetically pleasing visual layout.
  - Fruchterman-Reingold with Annotation (prefuse beta) – Visualization which lays out nodes based on some force between them.

  - DrL (VxOrd) – A force-directed graph layout toolbox focused on real-world large-scale graphs.
  - Specified (prefuse beta) – Visualization tool for use with graphs having pre-specified node coordinates.

  - Circular Hierarchy – Generates a circular visualization of the output produced by a multi-level aggregation method such as Blondel Community Detection. Result is a Postscript file.
  - Bipartite Network Graph - Generates a bipartite network visualization of the output produced by Extract Bipartite Network algorithm.