Network Analysis in Gephi

Alex Leslie

What is Network Analysis?

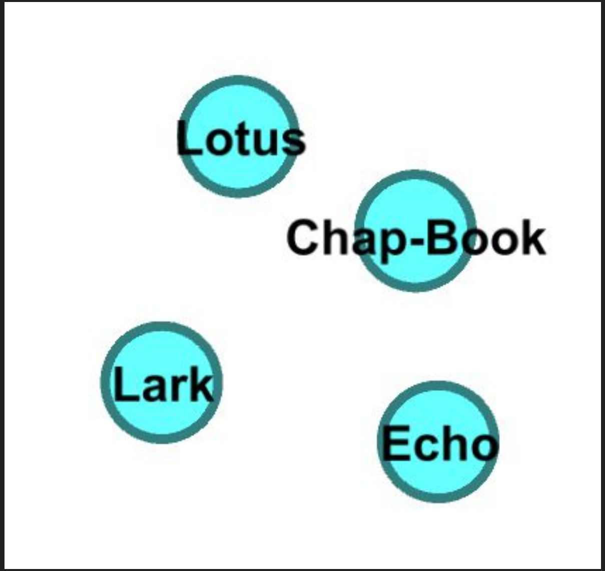
What is Network Analysis?

Network analysis is the study of the system of relations between entities.

Rather than emphasizing entities in themselves, single relations, or a group as a unit.

Sociologists tend to emphasize agents and social relations, but that needn’t be the case: any system of relations between entities can be represented as a network, provided the logic of relation is made clear.

Nodes

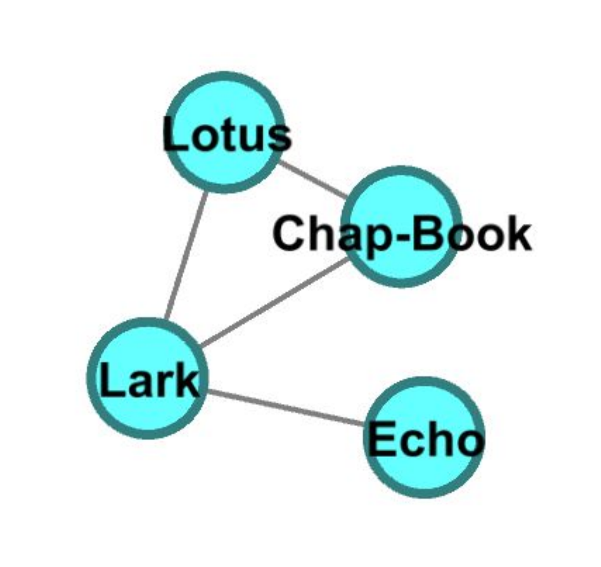


Networks are represented, both visually and computationally, as nodes and edges.

Nodes are the entities within a network, the subjects of the relation.

Any kind of entity can be represented as a node. In this example nodes are literary magazines, but they might just as easily be people (as agents or references), places, or any other objects.

Edges

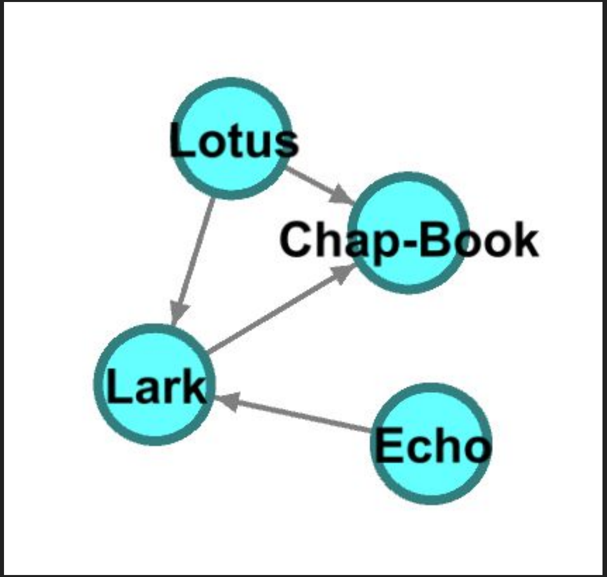


Edges, also sometimes called links, are the relations in the network.

Any kind of relation can be represented as an edge. In this example edges are citations, but in other networks an edge could represent a categorical relation (like friendship or shared mode of production) or a quantifiable relation (like correspondence between individuals or co-appearance in the same program / text).

If the relation is quantifiable, edges are weighted. Gephi weights edges by thickness and proximity. Highly-weighted edges are called strong ties; lightly-weighted edges weak ties.

Directed vs Undirected



There are two kinds of edge to represent two different kinds of relation: undirected and directed.

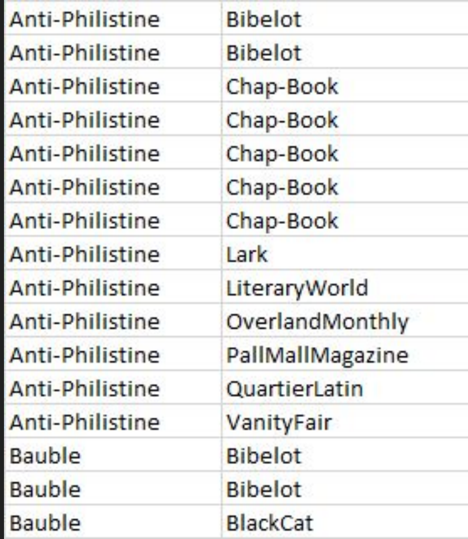
An undirected edge represents a relation without hierarchy or differentiation, like two composers appearing on the same program.

A directed edge represents a relation in which one entity occupies a different position than the other, like a letter recipient and its sender.

In many cases, one might use undirected edges to represent relations that are technically directed, depending on the nature of the data or the inquiry.

Assembling Network Data

Tidy Data



Anti-Philistine	Bibelot
Anti-Philistine	Bibelot
Anti-Philistine	Chap-Book
Anti-Philistine	Chap-Book
Anti-Philistine	Chap-Book
Anti-Philistine	Chap-Book
Anti-Philistine	Chap-Book
Anti-Philistine	Lark
Anti-Philistine	LiteraryWorld
Anti-Philistine	OverlandMonthly
Anti-Philistine	PallMallMagazine
Anti-Philistine	QuartierLatin
Anti-Philistine	VanityFair
Bauble	Bibelot
Bauble	Bibelot
Bauble	BlackCat

The basic format at the heart of all network data is an edges list.

This is simply a spreadsheet with two columns where each row is an edge connecting two nodes (the contents of each cell).

Gephi can work with several different file formats. The most universal, however, is a .csv file, so I recommend using this. It can still be opened in Microsoft Excel or Google Sheets; simply change the format option in the “Save As” menu.

Repeat edges need to be recorded either as duplicated rows or as a total in a third column.

Nodes.csv and Edges.csv

If there is additional data about the nodes and edges in a dataset pertinent to analysis of the network - distinguishing the kind of entity or manner of relation - it is necessary to keep separate spreadsheets for each. (You can always begin with just an edges list and add a nodes spreadsheet or additional columns later.)

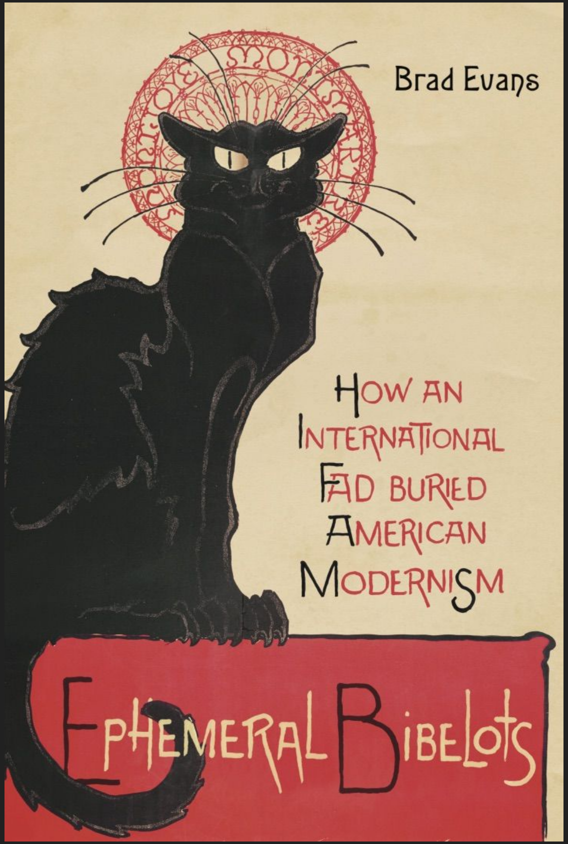
Gephi can be a bit picky about vocabularies

- Nodes spreadsheet
  - Node name must be in a column titled “ID”
  - If labels are other than ID, include as a separate column titled “Labels”
- Edges spreadsheet
  - First column, containing the first node of an edge pair, must be titled “Source”
  - Second column, containing the second node of an edge pair, must be titled “Target”
  - If using a separate column for weights, must be titled “Weight”

Sample Data: Evans' Bibelots Index

The sample data used in this workshop is lightly adapted from Brad Evans' (Rutgers English) dataset of ephemeral bibelots, available here: https://sites.rutgers.edu/bibelots/

The data was gathered as part of *Ephemeral Bibelots: How an International Fad Buried Modernism* (2019) to track the aesthetic, social, and citational relations among an international group of ephemeral, short-lived, faddish, cliquish, Paris-looking, avant-garde magazines in the 1890s.



Visualizing the Network

A Basic Undirected Network in Gephi

- Open Gephi and start a new project.
- File -> Open, selecting “evans_bibelots_weighted.csv”
  - The first window should recognize this file as an “Edges table”; if there were no “Weight” column or column headings, Gephi would recognize it as an “Adjacency list” (or edges list). Click next.
  - The second window will ask about the data types of any additional columns. Change “Weight” to “Integer” and click finish.
- The next window will be an import report; if there are any problems with the data, they’ll show up here. Change the graph type from Directed to Undirected for now and select “Append to existing workspace.”
- Gephi will plot everything randomly.

A Basic Undirected Network in Gephi

- In the upper left, switch from Overview to Data Laboratory.
    - In the Data Table, you can toggle between Nodes and Edges. Gephi has automatically populated the former from the later, but there’s one thing missing: labels. On the bottom bar, click “Copy data to other column” and select ID; the select that you want to copy these values to the Label column.
- Go back to Overview. Select the drop-down menu in the Layout tab in the lower half of the left-side panel. These are different algorithms for the layout of the network. I recommend Force Atlas, but feel free to try them out. You can also tweak the settings for each algorithm. Note that the same data never produces the exact same visualization, even though it is usually close.
- Scrolling zooms; right-click and hold to pan

Aesthetics

The border around the network visualization includes several options:

- Click the black T in the lower left to add labels; in the lower middle you can change size or font.
- The sliding scale to the left changes the relative thickness of edges
- The upper left part of the border has options for free editing: adding nodes, edges, sizes
- I think the most useful of these is the paint bucket, which colors a node and its neighbors

The Appearance tab at the top of the left-side panel includes general settings:

- Click the color palette icon to change node colors.
- Click the concentric circles icon to change node size.

Mess things around as much as you want; we’re going to scrap this first network and load a more data-rich version to explore more features.

A Directed Network with Categories and Dates

- File -> Open, selecting “evans_bibelots_nodes.csv”
    - Gephi should recognize this as a nodes table. Click next, finish, and ok to create a new workspace.
    - Switch to Data Table, click “Copy data to other column,” select ID, then select that you want to copy these values to the Label column.
- File -> Open again, selecting “evans_bibelots_edges.csv”
    - Gephi should recognize this as an edges table. Click next and finish. The third prompt should recognize this as a directed graph in the drop-down window. Instead of creating a new workspace, however, select the radio button to append this data to the existing workspace before clicking Ok.
- Select the “T” in the lower left to add labels, change the label size as desired, and run Force Atlas. This gets us pretty much where we were.
- The arrows represent the directionality of edges; their size can be adjusted with the sliding scale next to the “T” in the lower left.

A Directed Network with Categories and Dates

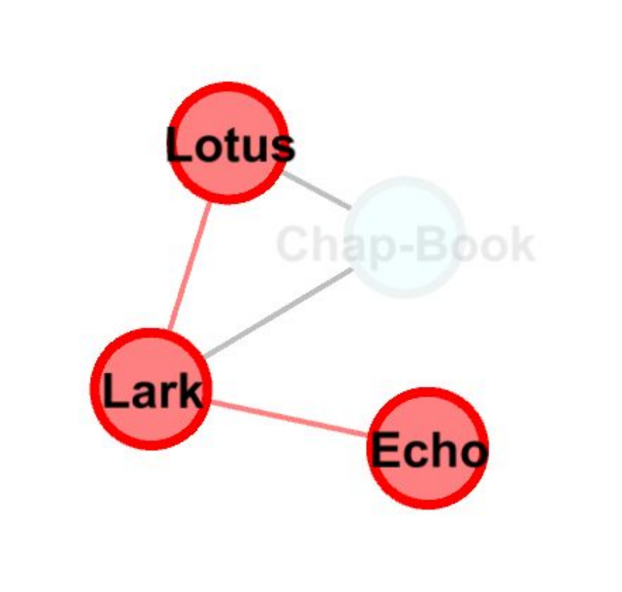
- A separate file for nodes allows us to include additional categorical data for each node.
- This dataset wasn’t created by looking at every magazine for which there is a node: it was compiled from a few dozen magazines, many of which reference other, unconsulted magazines.
- This means there are two types of nodes, which I’ve categorized as internal and other. To visualize this difference, click Nodes in the Appearance window, then the color palette icon, then Partition; finally, select type (the name of the column in our data) in the Choose an attribute drop-down menu.
- You might record types for each edge in order to similarly visualize categorical differences between edges as well. For this data, a column recording whether the citation was positive or negative would be instructive.

A Directed Network with Categories and Dates

- With a directed network, there are two additional options for sizing nodes. Select the concentric circles icon and then Ranking. In the Choose an attribute menu
- In-Degree corresponds to the number of edges targeting a node; Out-Degree corresponds to the number of edges for which that node is the source. In this network, then, In-Degree represents the number of times a magazine is cited.
    - In this network, out-degree represents referentiality: it tells us about a magazine's practices (aesthetic and/or social) of citationality. In-degree, by contrast, represents something like more like clout: it tells us about the extent of a magazine's reputation.
- We can adjust the text size in the same way: click the two Ts, then Ranking, select In-Degree, and adjust the min/max sizes as desired.

Analyzing the Network

Distance, Degree, Eccentricity

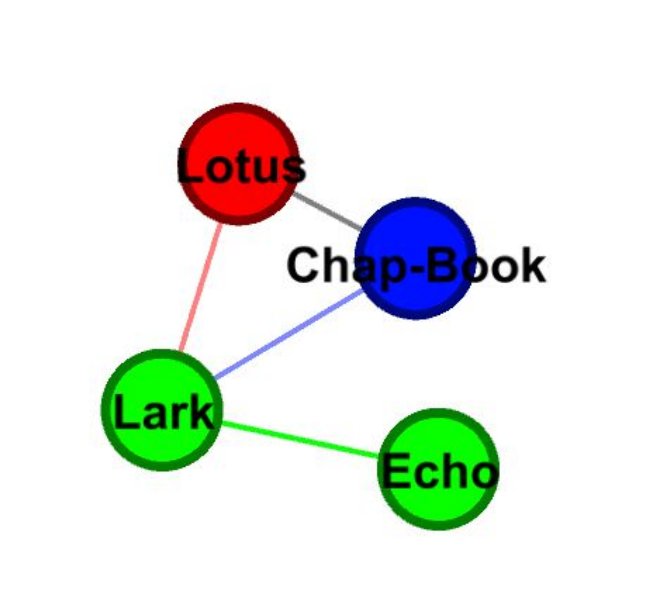


There are several ways to measure the connectivity of any given node in the network. Most rely on the concept of distance: the number of edges in the path between any given two nodes.

Degree is simply the number of edges a node shares.

Eccentricity is the distance between any given node and the most distant node from it. Lower values indicate more direct connectivity: the Lark’s eccentricity here is 1, all others are 2.

Closeness, Betweenness

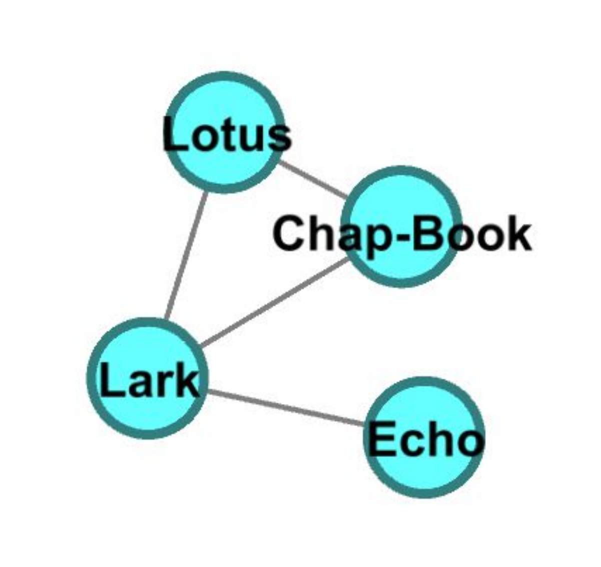


Closeness centrality is the average distance between any given node and each other node. This is calculated as: # of paths / sum of distances. The Lark is 3/3=1, the Lotus is 3/4=.75. Confusingly, you will sometimes see the equation flipped.

Betweenness is the number of times any given node appears on the shortest paths between all other nodes. Here the Lark has a betweenness of 2 while the others have 0.

High scores for degree, closeness, and betweenness indicates a highly-connected node that plays an important part in sustaining the network.

Diameter, Average Distance



We can also describe the network as a whole.

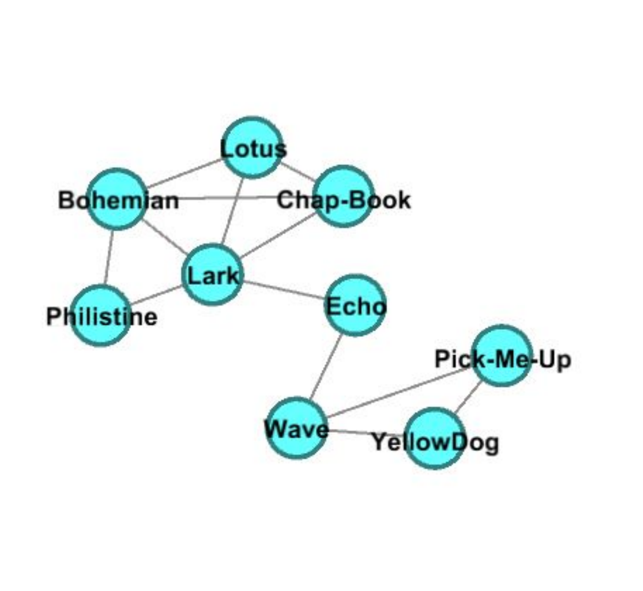
Diameter is the distance between the two most distant nodes.

Average distance is the average of all distances between all possible node pairs (here, 8/6=1.33).

Density is the proportion of the number of edges to the number of maximum possible edges (here, 4/6=.66), where a density of 1 would indicate a complete, or maximally connected, network.

The distribution of values of a particular metric for each node - for example, closeness centrality - can also be used to describe a network.

Clusters, Components, Bridges



Networks can develop subgroups in several ways.

Clustering coefficient measures the proportion of a node’s direct neighbors that share edges too. Here the Lotus has a clustering coefficient of 1; the Lark 4 edges among neighbors / 10 possible edges = .4.

Strongly connected components are groups with high density, small average distance, etc. among themselves even when the larger network doesn’t. Here there are two.

A node connecting components is called a bridge. Bridges have a relative paucity of local connections - low degree, low clustering, or high eccentricity - but high betweenness or centrality. Here, the Echo.

Filters

Gephi includes several filters to narrow down a larger network; these are found on the right-side panel under the Filters tab. Options include:

- Filter by attributes of nodes
  - Partition, Range, Inter-Edges
- Filter by edge strength
  - Edge weight, Edge type, Mutual edge
- Filter by topology, or position in the network
  - Degree range, Mutual degree range, In-degree / Out-degree range

When building your dataset in practice, it is usually preferable to include all data in one file and filter later rather than creating many smaller, pre-filtered files. This

Filters

To deploy a filter, simply drag and drop into the Queries window below and click Filter. Some filters will have sub-categories: for example, when filtering by Range there are a number of centrality measures to choose from. Most will require specifying the exact value by which to filter.

As a test, try the Topology filter “Mutual degree range.”

Gephi will always report the number of nodes and edges excluded by a filter in the Context section at the top of the right-side panel.

To combine filters, expand a filter in the Queries tab and drag the new one into it.

Click Stop to undo a filter; right-click it and select Remove to remove it.

Statistics

Gephi can calculate all of these metrics and more. On the right-side panel, select the Statistics tab and try some out.

Gephi will produce a report showing the distribution of values for each node for the network as a whole, which can be printed.

Measures of the overall network will display in the right-side panel itself.

To see metrics for individual nodes, select the arrow-with-a-question-mark tool in the left-side toolbar and then select any node. This will open a new tab in the left-side panel.

When (Not) to Network

Does the visualization add something that descriptive analysis could not efficiently?

Is the pertinent analysis about a system of relations rather than a number of entities or individual relations?

Does the concept of a network undermine the nature of the relations in question?

Can visualization or computation point to areas of further research, even if a network isn’t pertinent itself?