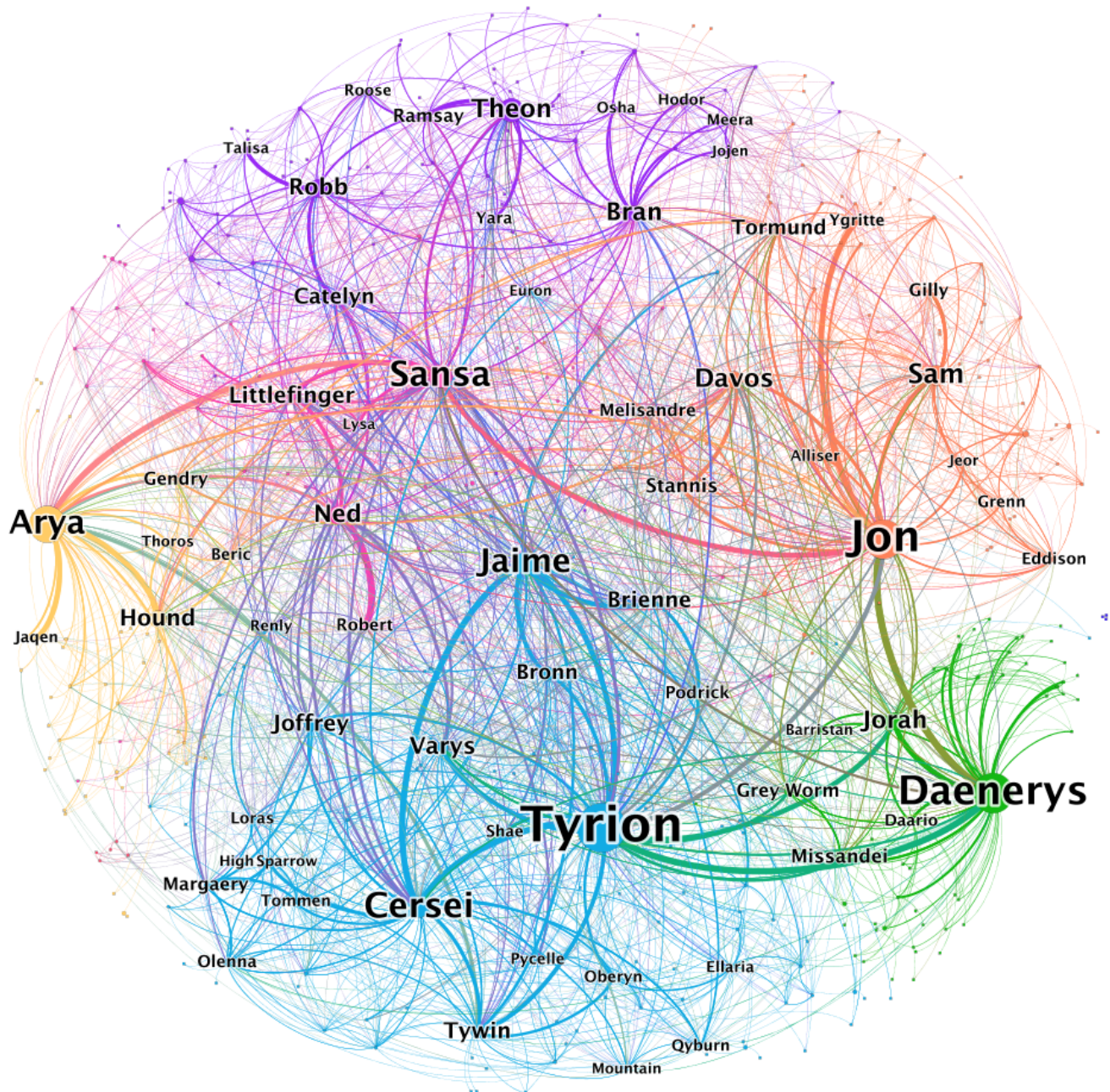


Introduction to Network Analysis

A Network Analysis is simply a study of relationships among a group of connected things. The "things" could be people, cities, cells in the body, etc.! The idea of a network analysis is to visualize these relationships, so you can see the patterns (or lack thereof) in connections within the network.



We use network analysis to ask questions about the players in the network, like

- Who is the most/least influential?
- What are some of the roles of the different things in this network?
- How connected are the things in this network?
- Are there distinct subgroups or "neighborhoods" in this network?

Objectives:

- Understand how network analyses help answer research questions with quantitative and visual components.
- Identify parts of a network
- Analyze relational data using different modes of centrality to determine the influence of individual nodes within the network
- Clean and model data according to a specific question
- Use Palladio to customize and interrogate the network
- Observe the existing relationships, gaps, and potential biases in the network

Lessons:

1. [Network Analysis Terms](#)
2. [Analyzing Relational Data](#)
3. [Data Modeling for Visualizing a Network](#)
4. [Software for visualizing a network analysis](#)

Required Readings:

- Heather Froehlich, "A Gentle Introduction to Excel and Spreadsheets for Humanities People," 2021.
<https://hfroehli.ch/2021/06/17/a-gentle-introduction-to-excel-and-spreadsheets-for-humanities-people/>
- Rawson and Muñoz, "Against Cleaning." 2016. *Curating Menus*,
<http://curatingmenus.org/articles/against-cleaning/>.
- Scott Weingart, 2011. "Demystifying Networks," *Journal of Digital Humanities* 1(1).
<http://journalofdigitalhumanities.org/1-1/demystifying-networks-by-scott-weingart/>

Further Readings and tutorials (optional)

- Thomas Padilla and Brandon Locke. "Introduction to Network Analysis."
<http://www.thomaspadilla.org/cytoscape/>
- Miriam Posner, "Network Analysis."
<http://miriamposner.com/classes/dh101f16/tutorials-guides/data-visualization/network-analysis/>
- Miriam Posner, "Creating a Network Graph with Gephi."
<http://miriamposner.com/dh101f14/wp-content/uploads/2014/11/Creating-a-Network-Graph-with-Gephi.pdf>

- Katayoun Torabi. Introduction to Gephi. 2020 Programming4Humanists Presentation
- Programming Historian, "From Hermeneutics to Data to Networks: Data Extraction and Network Visualization of Historical Sources." <https://doi.org/10.46430/phen0044>

Ethical Considerations:

Network analysis visualizations are excellent ways to show what's going on in a complex network, but they cannot stand alone! Even the most clearly labeled network needs context. When you are creating a networked graph, include some writing that gives context to the network, and shows the things you chose to highlight. Then include that explanation alongside the graph in an accessible way.

Another thing to consider when making highly visual graphs is how accessible your visualization may or may not be to everyone, including those with visual impairments. Consider the following when designing a graph:

- Are your labels large enough to read?
- Are the colors **in high enough contrast** to distinguish from each other?
- Is there alternate text or description somewhere that explains the graph with as clearly worded text as possible?

Finally, data used in a network analysis is not free from the influence of human bias. When you look at a network, don't just look for patterns, but given the context of the network, see if you can address what's missing, or where there might be slants that favor certain players over others.

Acknowledgements

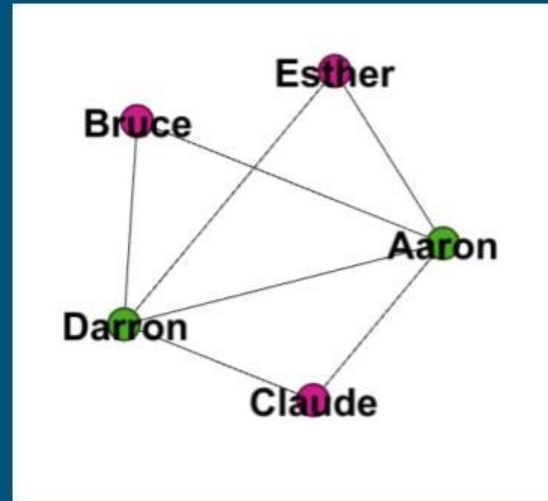
- Lecture notes from Dr. Katayoun Torabi, Programming for Humanists Instructor, Center of Digital Humanities Research at Texas A&M University.
<http://programming4humanists.tamu.edu/>
- Images of sample networks from
<https://www.coursera.org/lecture/networks-illustrated/closeness-centrality-part-ii-cqA9S>
- Images of steps in Palladio used from <https://hdlab.stanford.edu/palladio/tutorials/data/>.

1. Network Analysis Terms

Most software that facilitate Network Analyses use a different vocabulary. You'll learn about what each part of a network is in this section, and in the next section, we'll talk about different ways analyze those parts.

Networks

Network

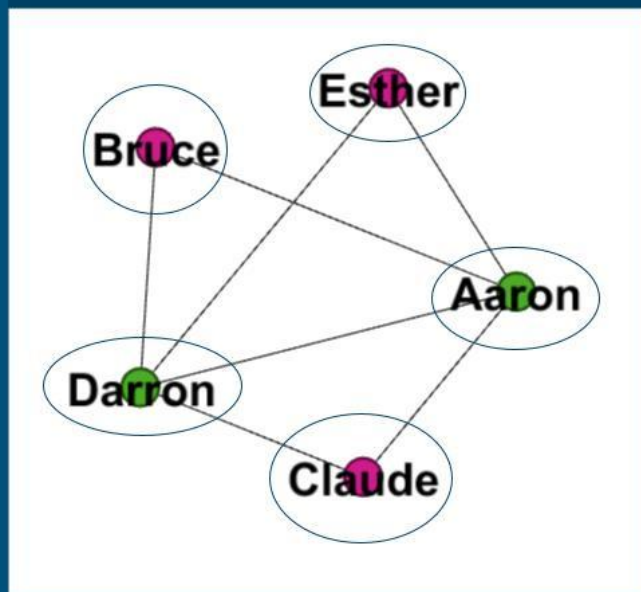


A network is a visual representation of relationships between entities. The network alone does not always share the context of the relationships (e.g. "These are people who talk to each other in this book," or "These are the flight paths among major airlines,"), but they do highlight relationships. In doing so, they show how different players dominate (or not!) those relationships.

Nodes

Node

An "entity"

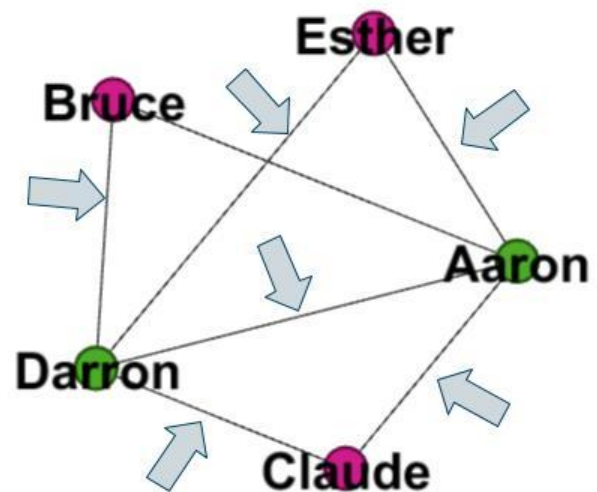


A node represents someone or something in the network. For the Smith Papers project, it will most likely be a person. Nodes can be different colors, depending on how influential they are, what subgroups within the network they belong to, etc.

Edges

Edge

A "relationship"

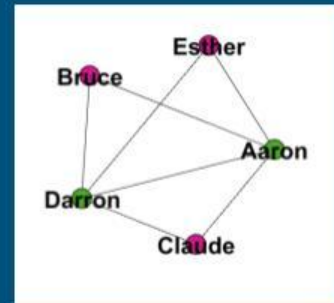
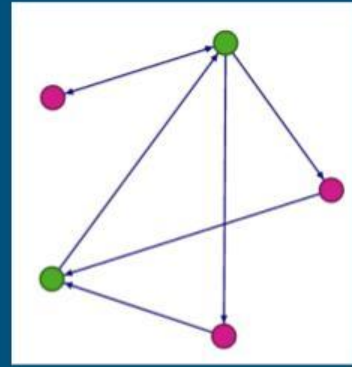


Edges represent the relationships between nodes. They usually indicate some kind of interaction in a relationship (e.g. conversation, travel, transactions, etc.), and depending on that interaction, an edge can be one of two kinds:

- Undirected - an undirected edge indicates that the interaction is reciprocal, like a two-way street, a friendship, a characteristic, etc.
- Directed - a directed edge indicates that the interaction is not reciprocal, like a one-way trip, a letter sent to someone, a payment for something, etc. In directed edges, there are those that give, and those that receive (just like in a *real* relationship!):
 - Source - the source is the originator of the interaction, or the giver. Sometimes this is called "out-degree."
 - Target - the target is where the interaction is directed, or the receiver. Sometimes this is called "in-degree."

Edges

- Directed edge - indicates a non-reciprocal relationship (e.g. letters sent, someone being spoken to, etc.)
- Undirected edge - indicates a reciprocal relationship (friendship, conversation, etc.)



Paths

Networks

- Can measure how connected the individuals are
 - Lengths of paths
 - Density of the entire network
- Can have clusters
 - proximity of individuals to each other

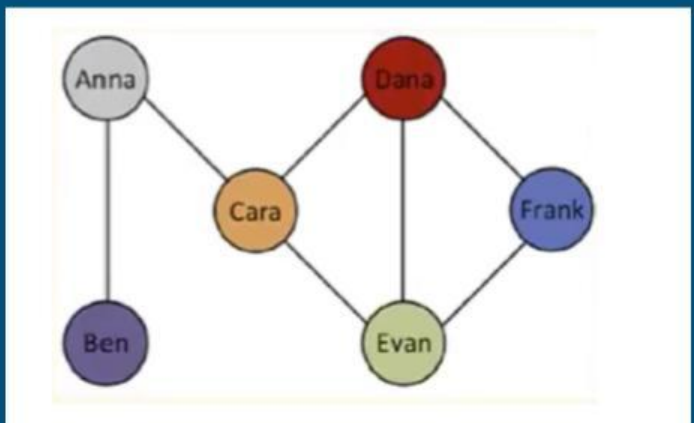


Image from <https://www.coursera.org/lecture/networks-illustrated/closeness-centrality-part-ii-cqA9S>

Paths are the length from one node to the next. In this image, the path length from Ben to Anna is 1, but the path length from Ben to Cara is 2, and so on. The average length of each path will tell us how

dense the network is, or in other words, how connected everyone is to each other. If everyone is connected to everyone by 1 path, the network is 100% dense!

...just a few more terms! Stick with me!...

Check-in

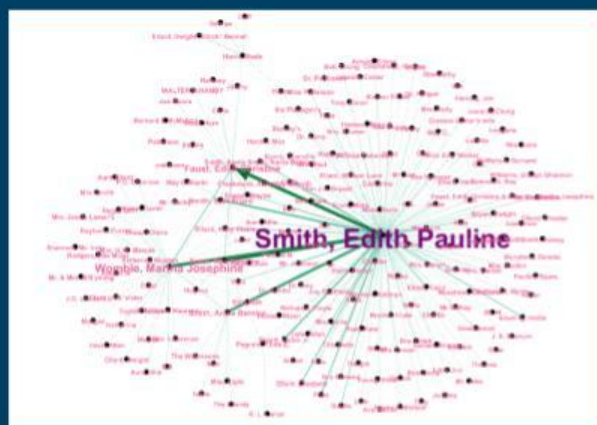
- What questions can a network analysis help answer?
- What roles do particular entities play within a network?
- How connected or related are individuals within a network?
- What do the groups or subgroups look like?
- True or false: You can have a network where nodes are not connected.
- Which kind of edge indicates that an interaction between two nodes is reciprocal?
- If all of the nodes in a network are connected to each other with an edge, how dense is the graph?

2. Analyzing relational data

Did that phrase make your stomach turn? Don't worry! We are not actually doing math in this class, but we will be using mathematic concepts. We analyze networks by calculating the ways nodes influence each other. Let's break it down.

Degree

How connected or influential an individual is within the overall network



Centrality

Centrality is the relative influence of individual nodes within the network. In the image above, the nodes have different sizes and colors to indicate their influence. Notice that Pauline has several direct connections, so they are the largest node. Other nodes, like Martha, have fewer connections, and therefore appear smaller. From this, we infer that Pauline has a more influential role in the network.

In this lesson, we will measure centrality with 3 different metrics:

- Degree
- Closeness
- Betweenness

Degree Centrality

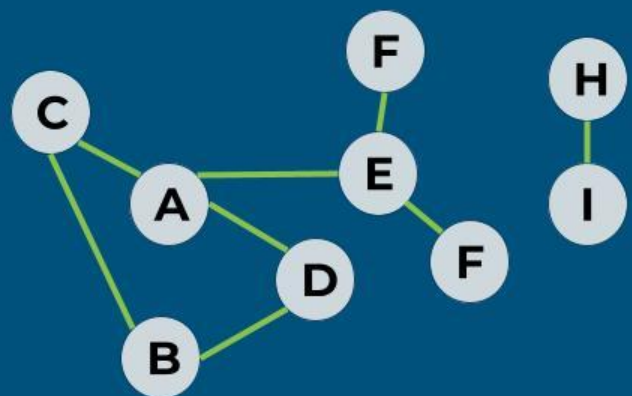
The degree of a node is how connected or influential a node is within the network, so degree centrality counts the number of direct connections a node has. In the previous image, Edith Pauline Smith has the highest degree centrality because she has the most immediate connections. To indicate this, we have made the size of node labels correspond to degree, which for Pauline, is the largest.

Closeness Centrality

Another simple metric is called closeness centrality, which measures which node has the shortest average path to the rest of the nodes in the network.

Centrality

- A. Degree Centrality - number of connections
- B. Closeness Centrality - closeness to the entire network (who has the shortest average path to everyone else?)



Unlike degree, which measures direct connections, closeness measures the proximity of a node to all the other nodes through average path. There are 8 total edges in the network above. If we wanted to calculate the closeness centrality of B, we'd count the shortest path from B to every other node, and take the average.

B -> A = 2

B -> C = 1

B -> D = 1

B -> E = 3

B -> F = 4

B -> G = 4

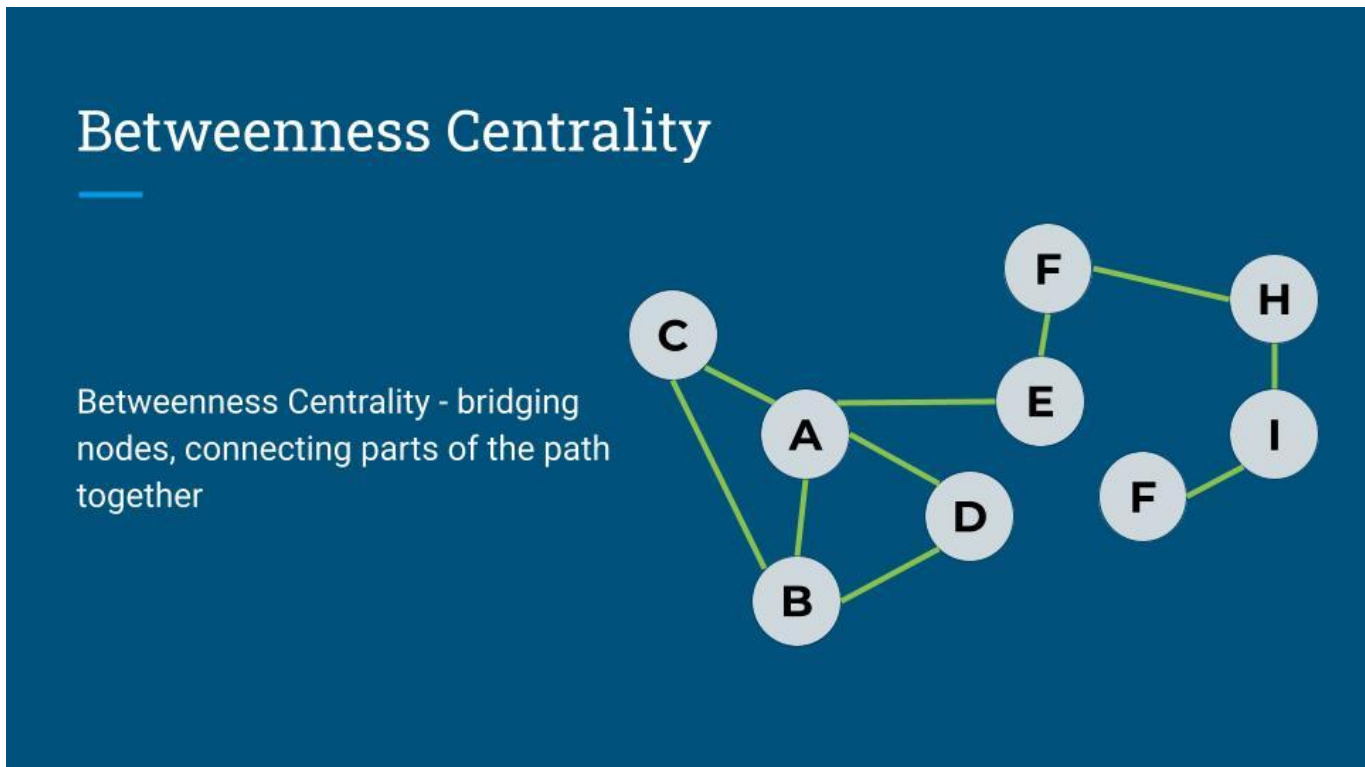
B -> H = 0 (there are no paths to H from B!)

B -> I = 0 (there are no paths to I from B!)

B has a closeness centrality of 1.875. (There are 15 total paths to every other node, and 8 total edges, so $15/8=1.875$). So this node has more direct connections (2), but is less "close" in terms of the entire network (1.875).

Betweenness Centrality

Degree and Closeness measure influence through connections and paths. Betweenness centrality indicates nodes that are "bridges" between different groups in a network. In other words, they glue pieces of a less dense network together, or they serve as a "go-between" for 2 distinct parts of the network.



Check-in

- How would you define centrality in your own words?
- How would you define degree centrality in your own words?
- How would you define betweenness centrality in your own words?
- Draw a simple network with pen and paper that shows nodes with high degree centrality.

3. Data Modeling for Visualizing a Network

In order to tell the software (which we go over in the next section) how to recognize which nodes are connected, and by what paths, we are going to create a spreadsheet with "tidy data." Each column will have a variable, and each row will be an observation of that variable.

The most important thing to remember when structuring data for a network analysis is this: what question are you trying to answer with a network analysis graph? Are you trying to see who someone is writing to in a letter? Are you trying to find out who has written the most letters and who has received the most? Are you trying to figure out who gets mentioned more times by one person than

another? All of this matters. For instance, if we wanted to see where the letters were coming from and who is writing to whom, we could structure the data like this:

Source	Target	Address	Return Address
Pauline Smith	Sam H. Smith	Pittsboro	Jackson
Christine Smith	Pauline Smith	Starkville	Pittsboro
Pauline Smith	Martha Smith	Pittsboro	Meridian

In this table, each column heading (the variables) has one observation (the information in one letter). That's tidy data!

Say, for instance, though, that we want to see all of the people mentioned in several letters to get a better idea of the network of people in the lives of the Smith family members and their friends. We could structure our data so that each source (the letter writer) has a target (the person they write to and the people they mention), and each target has either a reciprocal (undirected) or non-reciprocal (directed) edge.

Check in

Restate the research question for the data set at hand, and sketch the model of the data on paper. Example: If we want to see all of the people mentioned in several letters to get a better idea of the network (for the sample data, the network = people in the lives of the Smith family members and their friends), what would our columns and rows look like? What would our data model look like?

For sample data set: We could structure our data so that each source (the letter writer) has a target (the person they write to and the people they mention), and each target has either a reciprocal (undirected) or non-reciprocal (directed) edge. What would this look like in a spreadsheet? Sketch this out in pen and paper.

Cleaning Data

Wouldn't it be nice if you could just copy and paste some names in a list into a spreadsheet, and click a button to make a nice network graph? IT NEVER WORKS THAT WAY! But that's part of the work of the digital scholar - is to structure data through a process of cleaning and testing it, in order to see if they can answer a question through visualization. Hadley Wickham, author of ["Tidy Data,"](#) says:

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson 2003). Data preparation is not just a first step, but must be repeated many times over the course of analysis as new problems come to light or new data is collected.

For the names in our collection of letters, we have divided the data into two groups, unstructured and structured:

- **Name tags - Unstructured.** This is what it looks like when you export the metadata tags you created from Transkribus, and copy and paste them in a sheet together. You will be adding to this spreadsheet with letters that you transcribed.
- **Name tags - Structured.** This is what it looks like when you have added another layer of structure to those tags.

Even though these names are structured, we're not done! We have to check the data for things like inconsistencies, duplicates, and errors! There are a few ways to do that.

1. Copy and paste the names of people in the tags as transposed data into the ongoing "Name Tags - Unstructured" sheet, so we have a list of all letters and those who are named. ([Watch this in a video here.](#)):
 - Highlight the column of names in the sheet of exported tags and copy them
 - In the Unstructured tags sheet, enter the sender and recipient, then paste the names in the first column under person 1. This will look strange at first! But we haven't transposed them yet...
 - Cut this list you just pasted, and then paste them again using the Paste Special > Paste Transposed option. Ta da!
2. Clean the names that you can ([Watch this in a video here.](#)):
 - De-duplicate names that appear more than once, since you just want one representation of that person per letter. (To measure our network, we are not including each time a person is mentioned in a letter...just that they are mentioned at all!)
 - Replace *known* names with their authority control names (Last name, First name) or reconcile names that are slightly different (i.e. Bro Breland and Brother Breland) for consistency. We have identified some people by first name and context, and listed them in [this CSV](#).
 - If you don't know them, that's ok! Remember, we aren't trying to erase anyone's name by assuming (see "[Against Cleaning](#)" again!). Kate and Katherine might not be the same person. We don't know! Embrace the chaos and keep assumptions to a minimum!

Bowling, Hattie
Boyd, Minnie Claire
Brewer
Bro. Brealand
Bro. Breland
Bro. Brumer
Bro. Clark
Bro. Jones
Bro. Patterson
Bro 'McKay'
Bro Breland
Bro Clark
Brother Clark
Brother Clark
Broyles boy
Bryan, Dwight

This process will take a while, but remember, that is normal! Give yourself time, and take plenty of breaks!

Check-in

Write a detailed list of things that you did to clean up the data you exported. Did you de-duplicate names? Did you make assumptions for similar names? If so, what assumptions did you make? Write these down as precisely as you can so that if someone else were to follow your directions, they'd get similar results. Think of this like a recipe with explicit 1-2-3 instructions, like 1. De-duplicate column C, 2. Sort column C to find similar names, 3. Names changed: Foss = Jack Foster, etc.

4. Software for visualizing a network analysis

We have prepared our data (it won't be the last time!), so now it's time to see what that looks like in a network analysis software's visualization. There are several different softwares for visualizing a network, and they all serve different purposes. We are going to use [Palladio](#) because it is free, has a relatively low accessibility bar, will work in your browser, and most importantly, will allow us to use the visualization to answer our questions: What does the network of people in the lives of the Smith family members look like? Where are the connections and who facilitates them?

Preparing your visualization

1. Load your data

You can either copy and paste the cells from your spreadsheet (columnn headings included), or you can upload your file as a .csv file - *.csv, or comma separated value files are more flexible among different platforms, and is a simpler, less formatted version of tabular (or structured) data.*

Palladio will not support an .xlsx file!

Copying your data will look like this:

SmithPapers_Edges_1929-1939 ☆ ⓘ

File Edit View Insert Format Data Tools Add-ons Help *Last edit was made on April 4 by Hong Gao*

100% \$ % .0 .00 123 Calibri 11 B I A ↵

C1:E1282 fx Source

	A	B	C	D
1	Row	Letter	Source	Target
1263	100	mcj-dp018-19381212-smith-pauline	Smith, Edith Pauline and Ma	Smith, Edith Pauline
1264	101	mcj-dp018-19381110-smith-pauline-001	Smith, Edith Pauline	Faust, Edith Christine
1265	101	mcj-dp018-19381110-smith-pauline-002	Smith, Edith Pauline	Jack Foster Ellard
1266	101	mcj-dp018-19381110-smith-pauline-003	Smith, Edith Pauline	Dr. Sawphil
1267	101	mcj-dp018-19381110-smith-pauline-004	Smith, Edith Pauline	Charley's
1268	101	mcj-dp018-19381110-smith-pauline-005	Smith, Edith Pauline	Judge Kelly
1269	101	mcj-dp018-19381110-smith-pauline-006	Smith, Edith Pauline	Sam Hawkins Smith
1270	101	mcj-dp018-19381110-smith-pauline-007	Smith, Edith Pauline	Chester
1271	101	mcj-dp018-19381110-smith-pauline-008	Smith, Edith Pauline	Bob Young
1272	101	mcj-dp018-19381110-smith-pauline-009	Smith, Edith Pauline	Dr. Rudner
1273	101	mcj-dp018-19381110-smith-pauline-010	Smith, Edith Pauline	Philpot
1274	101	mcj-dp018-19381110-smith-pauline-011	Smith, Edith Pauline	Martha
1275	101	mcj-dp018-19381110-smith-pauline-012	Smith, Edith Pauline	Wayne's
1276	101	mcj-dp018-19381110-smith-pauline-013	Smith, Edith Pauline	Bernice
1277	101	mcj-dp018-19381110-smith-pauline-014	Smith, Edith Pauline	Sam Smith Ellard
1278	101	mcj-dp018-19381110-smith-pauline-015	Smith, Edith Pauline	Mrs Ayden
1279	101	mcj-dp018-19381110-smith-pauline-016	Smith, Edith Pauline	Katherine G,
1280	101	mcj-dp018-19381110-smith-pauline-017	Smith, Edith Pauline	Womble, Martha Josephine
1281	101	mcj-dp018-19381110-smith-pauline-018	Smith, Edith Pauline	Tom
1282	101	mcj-dp018-19381110-smith-pauline-019	Smith, Edith Pauline	Jenel
1283				
1284				

+ Sheet1

Cut Ctrl+X

Copy Ctrl+C

Paste Ctrl+V

Paste special ▶

Insert 1282 rows

Insert 3 columns

Insert cells ▶

Delete rows 1 - 1282

Delete columns C - E

Delete cells ▶

Sort range

Randomize range

Convert to links

Remove link

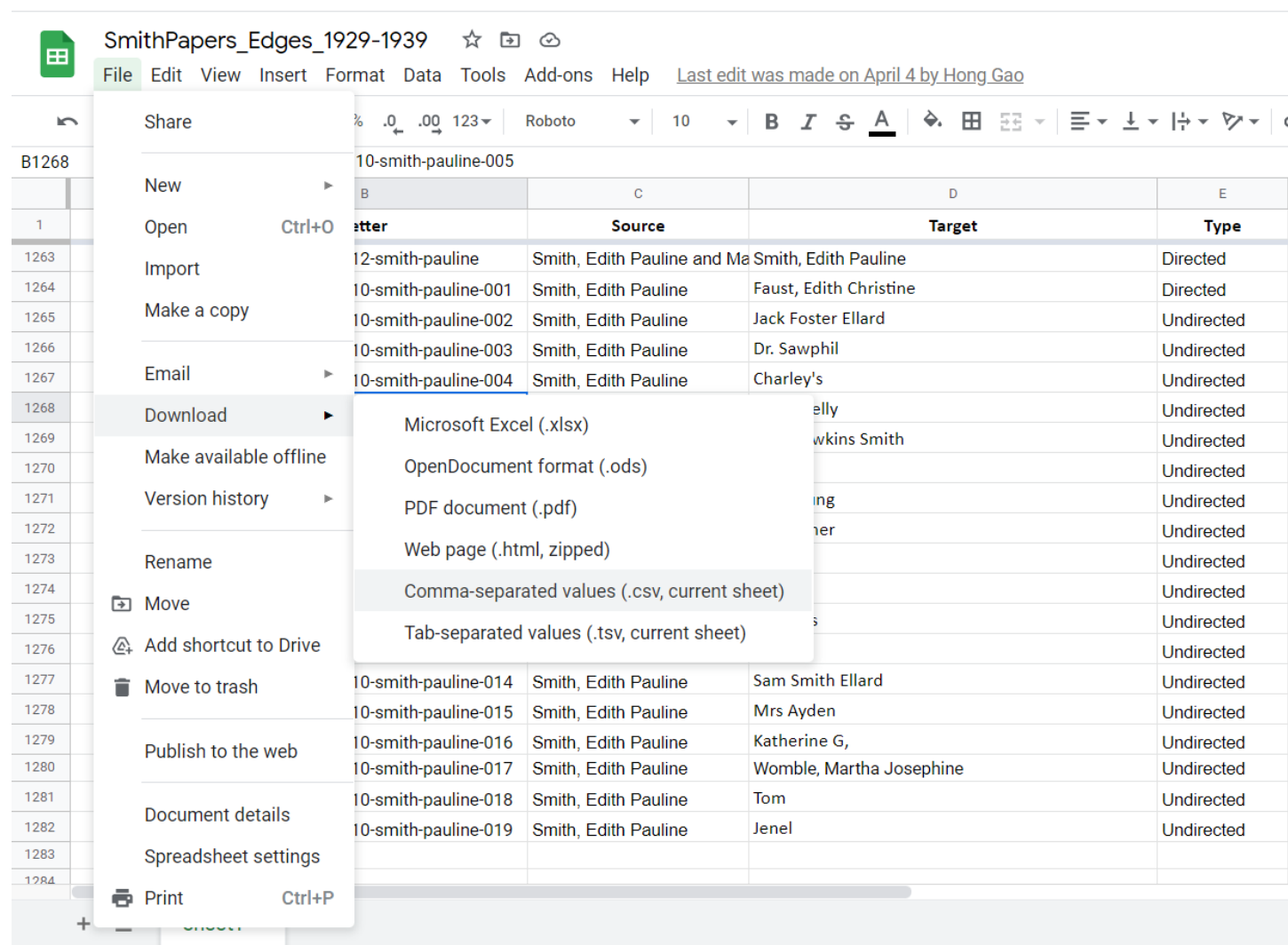
Insert link Ctrl+K

Open links Alt+Enter

Get link to this range

Define named range

Saving your data to upload will look like this:



SmithPapers_Edges_1929-1939

File Edit View Insert Format Data Tools Add-ons Help Last edit was made on April 4 by Hong Gao

Share

New

Open Ctrl+O

Import

Make a copy

Email

Download

Make available offline

Version history

Rename

Move

Add shortcut to Drive

Move to trash

Publish to the web

Document details

Spreadsheet settings

Print Ctrl+P

Letter	Source	Target	Type
12-smith-pauline	Smith, Edith Pauline and Ma	Smith, Edith Pauline	Directed
10-smith-pauline-001	Smith, Edith Pauline	Faust, Edith Christine	Directed
10-smith-pauline-002	Smith, Edith Pauline	Jack Foster Ellard	Undirected
10-smith-pauline-003	Smith, Edith Pauline	Dr. Sawphil	Undirected
10-smith-pauline-004	Smith, Edith Pauline	Charley's	Undirected
10-smith-pauline-005		elly	Undirected
10-smith-pauline-014	Smith, Edith Pauline	Sam Smith Ellard	Undirected
10-smith-pauline-015	Smith, Edith Pauline	Mrs Ayden	Undirected
10-smith-pauline-016	Smith, Edith Pauline	Katherine G,	Undirected
10-smith-pauline-017	Smith, Edith Pauline	Womble, Martha Josephine	Undirected
10-smith-pauline-018	Smith, Edith Pauline	Tom	Undirected
10-smith-pauline-019	Smith, Edith Pauline	Jenel	Undirected

2. Resolve flagged issues in Palladio

Palladio flags inconsistencies in data, like the use of special characters or multiple observations in one variable (i.e. more than one value in a cell--commas are tricky!). Use Palladio's verification tools, sorting options, and searching option to take another thorough look at your data.









Clicking on the little red dots will allow you to verify and look further at your data.

Provide a title to this project

Show details

Untitled

Primary table1281 rows

 Source	Text 
 Target	Text 
 Type	Text
 generated	Text
	

This is also an opportunity to evaluate if you need to go back and clean the original spreadsheet! For instance, if you are reviewing the issues, and you find that:

- a person's name recorded in 2 different ways,
- spelling errors, or
- the type of data (i.e. text, number, date, etc.) is displayed incorrectly,

you want to take this opportunity to clean your data again. This is part of the process!

Edit dimension

Title

Target

Data type

Text

Unique values

Search

Sort by Value

Bob M

Bob Young 3

Bob young

Bobby 2

Bobby Burns

609 values displayed. [Download](#)

Verify special characters

Multiple values

If the dimension contains multiple values, insert the delimiter string above

Extension

Choose a table

Add a new table

You can provide additional information about this dimension with data from another table.

Close

3. Choose which dimensions to visualize

Once you've verified the data and gone back to clean any leftover issues, it's time to visualize it. (Reality check: this isn't necessarily the end of cleaning data! A visualization can also make inconsistencies and cleaning issues apparent!). For the question of how to visualize the letter writer and their network of people (the person who received the letter and those mentioned within), click the **Graph** tab, then choose the following

- Source dimension - this is the sender of the letter
- Target dimension - these will be both the recipient and the people mentioned

4. Customize your graph

- Add facets to your data
- Size the nodes according to different dimensions (ours will just have one dimension - number of times they appear).
- Filter what appears in the network (Directed v. Undirected)
- Drag nodes manually to see different connections among the network. Who is connected to whom? Who bridges the network?

[Here is a brief video of what this process looks like.](#)

There are also Palladio's own Tutorials and FAQs for [loading data](#) and [customizing a graph](#). Both of which are helpful!

Check-in

Eyeball test - Look at the graph you were able to generate with Palladio. What questions can you start to answer with the image you see? What looks like it needs fixing?

An “Eyeball Test” of your visualization is the first step in looking for interesting features in your data. It can be a great way to start answering questions you posed when you collected the data, and can get you started asking more questions about interesting features in your graph.

Remember that data sets have context and are not without bias! Knowing what you do about this data set, imagine what is missing, or what is left out? Thinking also about your own preconceived notions, what did you notice first? How do you think collection of the material influenced how the data is visualized? (For the sample data set, the [contextual information about the collection](#) will help address this.)

Submit (optional): Submit a document with an image file (or several image files) of your network analysis in Palladio, and include a paragraph to explain what's going on in the graph, which should give:

1. a brief explanation of the collection of letters, and
2. a description what the nodes and edges represent within the greater network of the collection of letters, and
3. any insights you can provide on relationships or nodes to highlight (i.e. notes on who might have the highest degree centrality and why, or which node might serve as a bridge, etc.). If you need to zoom in on parts of the image to do this, include a zoomed-in screenshot of part of the network so that your explanation has a visual aid.