For my second blog post. I want to go over the original method I used to make the network graph.

1. DATASET ACQUISITION

Over 35 years of research, S.D. Goitein created over 35,000 index cards detailing his notes on the Cairo Geniza. Each index card has a description of a single fragment from the Cairo Geniza. Information from these index cards are the basis of the network graph I created. Here is an example of one {images/sd-index-card04-393-1.jpg}.

These notecards were then transcribed and put online by the people at the Princeton Geniza Lab. Thanks to them, we have excellent descriptions of tens of thousands of documents available online. The information available to me was each documents’ PGPID, a unique sequential ID; its written description; and the document type (letter | legal | ketubah | list | etc).

2. CLEANING THE DATASET

I first tried to get the data at scale with named entity recognition. However, standard named entity recognition was not trained on patronymic-based names. The names would be something like “Nahray ben Nissim” and the named entity recognition tools would recognize “Ben Nissim” or “Nahray ben” without understanding that its all part of one name. I struggled (and failed) to create my own named entity recognition tool which would work for the names.

Luckily, S.D. Goitein tended to have a fairly consistent method of writing notes for letter documents. He would write that letters are "from" a sender and then "to" the recipient. This means that instead of writing a complex machine learning algorithm to try to determine the names within the documents, we can just take the documents and feed it through some some data cleaning regular expression statements. All you need to do is take everything after the word “from” and before the word “to” as the sender and everything from the word “to” to a punctionation mark as the receiver, right? (SHOW EXAMPLE OF THIS WORKING)

Unfortunately, its not so simple (FIND A GOOD EXAMPLE AND PUT IT HERE). While this massively simplifies the operation, it means that there is substantially more data cleaning by hand (I'll get to that in a moment) and a massive loss of potential data.

Take for example, this letter:

Letter from Faraḥ b. Yosef, in Alexandria, to Yehuda b. Menashshe, in Fustat. Dating: May 12, 1069. Regarding shipments of gold and exchange of dinars for dirhams. Mentions the ships that sailed from Alexandria to the west, probably the first ones after the winter. The writer asks the addressee to drop a debt of 40 dinars that he borrowed from Abu Naser Thibt b. Sadan al-Bagdadi, a person who came to Alexandria in order to travel to the Maghreb, but missed the ship and returned to Fustat. Also Mentions details about the situation in Sicily and the battles there, especially the rise to power of the last Muslim ruler of Palermo, Ibn al-Baʿbāʿ. (Information from Gil, Kingdom, Vol. 3, #520) VMR

An algorithm which only gets the text from after "from" to "to" returns "Farah b Yosef, in Alexandria"

An additional algorithm which only gets the text from after "from" until the subsequent punctuation mark (with the exception of the "b." in Yehuda's name) would return "Yehuda b. Menashshe, in Fustat".

Basically, I got to this point and then cleaned it by hand by scanning through row by row for thousands of rows until all the names were consistent.

That was the easy part

What was much harder was making all the names consistent. For some reason, it seems that Goitein wrote different names for the same people at different times. For example, 5559 and 5601 have Musa b. Yishak b. Hisdai (5559) and Musa b. Yishak b. Hisda (5601) write to Yosef b. Ya’aqov b. Awkal. However, Hisdai and Hisda would not appear as the same person without rigorous text cleaning. Similarly, while both documents use the same name for b. Awkal, there are many other documents with him where he is only noted as "Ibn Awkal" or “Abu l-Faraj Yusuf b. Yaaqov b. Awkal” While it is not so difficult to determine that “Yosef” and “Yusuf” are the same, it gets more difficult when you’re trying to determine whether two people with similar names but wildly different honorifics are the same. Its nearly impossible for the untrained to determine that XX ha-hesed and XX the cantor are the same or different unless they go into the documents and determine who is who. This took an incredible amount of time and I am certain that I missed a substantial number of people who were actually the same.

3. MAKING THE NETWORK GRAPH

But once I had gone through and gotten the names, normalized them to be a standard spelling of the name, and standardized which person would normally have which honorific, it was easy to build a network graph of all the people who sent letters from and to each other.

I hate web development. I am going to force myself to become better at it but so far, I find it miserable. With the aid of made a network graph in d3.js with the help of Ben Johnson at the Princeton Geniza Lab, I created a small website to host it. Later, I added a method to view the letters by clicking on the nodes representing each individual and the links connecting them.

4. COMMENTARY

However, such a simple method means we miss a lot of other good data and forces the question of what the goal of collect data and making a letter network should be?

The basic version that I created did exactly what is listed above. I got the name of the author and the name of the recipient and simply put them on a graph.

First, there is additional person mentioned in the document, "Abu Naser Thibt b. Sadan al-Bagdadi." He is not the sender of the letter nor the recipient, however he is known to both of them. If the purpose of the network is just to show who sent letters to who, he should obviously be excluded. However, that's not really the purpose of a letter network, is it? Ultimately, we want to know who knows who and what the community at the time would have been. If we want to show a interpersonal map of 11th century networks as expressed in the Cairo Geniza, he should be equally expressed as a character in the network graph. This leads us to the question of how he should be represented? It's ultimately a shortcoming of my visualization capabilities that I honestly don't know. I think the graph would be an untangleable mess if every connection on every document is shown but ultimately "an untangleable mess" is pretty accurate for these networks! There is not a great solution to this.

The other thing this document shows is the difficulty of these documents. In my more complete list of documents which collects all people, not just the sender and receiver, "Abu Naser Thibt b. Sadan al-Bagdadi" seems to appear only in this letter. This appears to be the only time Thibt b. Sadan is ever mentioned. All we will ever know of him is that he "came to Alexandria" with the intention to travel to the Maghreb but missed a ship and returned to Fustat, had 40 dinars which he lent out.

I was surprised that someone with capital like 40 dinars would not have been mentioned elsewhere. So I looked for the components of his name. There is no other "Thibt" in the text. However, there are twenty "Thabit(s)" and eleven "Sadan(s)." Going through them, I find a "Sadan b. Thabit ha-Levi," who is almost the same person as "Sadan b. Thabit Baghdadi," who is almost certainly the same guy as "Sadan B. Thabit Levi Baghdadi." However, "Sadan b. Thabit" seems to not have sent a letter until 1135, 66 years after the letter where Thabit b. Sadan had been mentioned. These guys clearly seem like different people. There is also the incredible fact that one of Sadan b. Thabit's letters was only identified by Goitein as Sadan b. Thabit based on Sadan b. Thabit's handwriting!

This kind of discovery is fundamentally why it has been so difficult to quickly get clean information out of the Cairo Geniza. There are tons of different ways that people address each other and lots of similar names. Sometimes, they have to be checked by hand. The most exciting moments are where I discover a new way that someone spelled a name and it turns out that there are two relatively common ways to spell the same name and after recompiling the documents, it turns out that several people as listed in the documents are actually the same people, but with different honorifics in different letters.