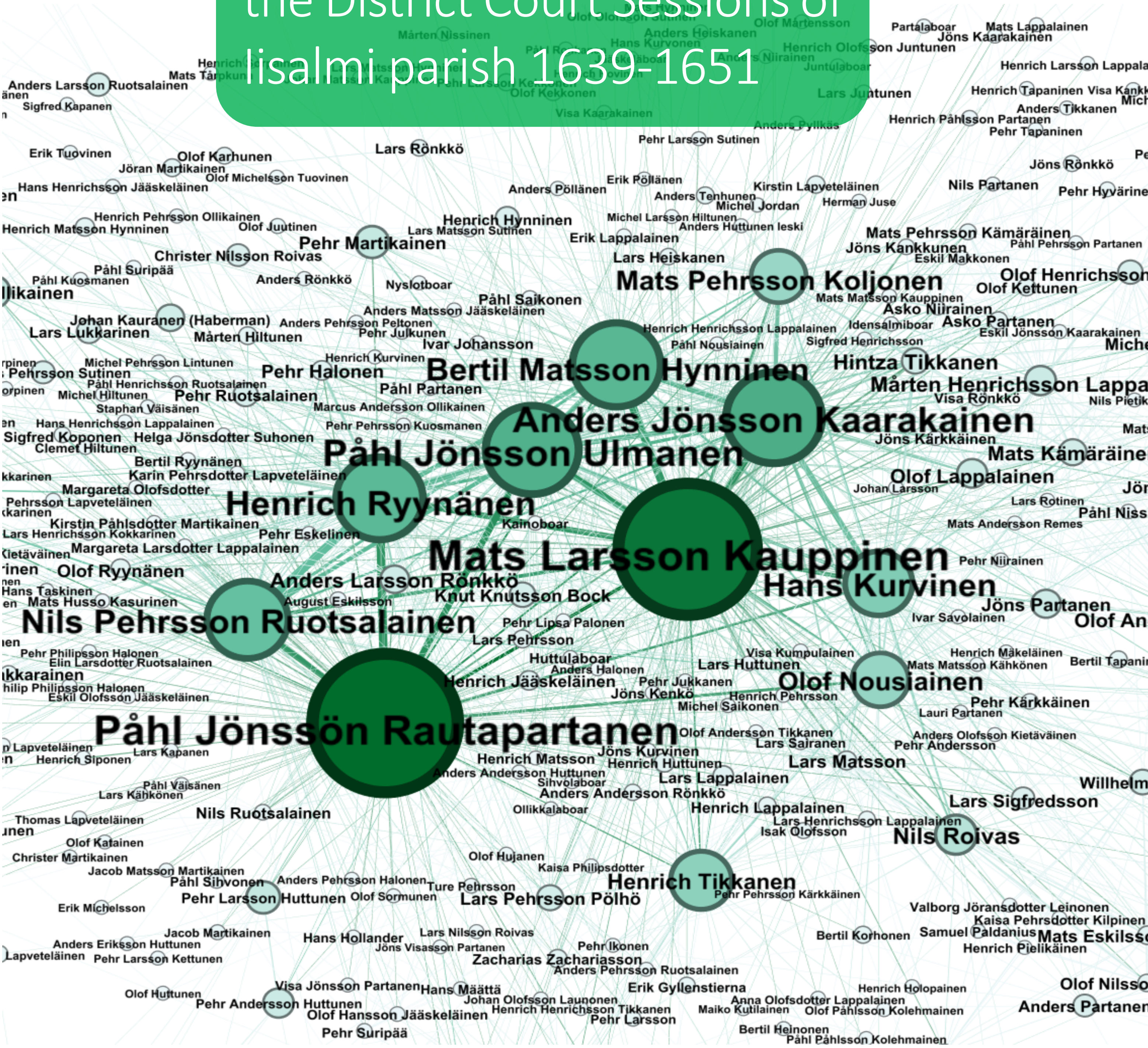
Social Network Analysis and the District Court Sessions of Iisalmi parish 1639-1651



Ville-Pekka Kääriäinen,

Researcher PhD Student at University of Helsinki,
Doctoral Programme in History and Cultural Heritage

Project work for the course “Computational literacy for the humanities and social sciences” (2020),
Faculty of Arts at the University of Helsinki

To see one version of the network visit: https://vilkaari.github.io/SNA_Iisalmi_1639-1651/network/

Contents

Humanities research questions

This course project was made to serve my PhD thesis. In my PhD thesis, I study the state-building process in the peripheral area of Iisalmi[1] (Pohjois-Savo, Northern Savonia) in the years 1639–1699. Traditionally the state-building process is examined from above and thus researchers have underlined the significance of macro-level institutions like the Riksdag of the Estates (Fi: *söötyvaltiopäivät*, Swe: *Riksdag*). In my research, I consider that the state-building process happened at the local level. My study is going to be a continuation to research that emphasizes the importance of personal agency as a key factor in the state-building process. This means that the state-building required activity at the local level by local officials (bailiff, sheriff, priest) and local people (peasants, lay members of the court). That is why it is important that I can perceive the local networks of interaction.

The relationships that I'm interested in:

- Relations between the crown officials and the common people (*rahvas*). For example, what was the relationship between the bailiff (*vouti*) and the peasants.
- Relations inside the peasant community. I'm especially interested to find out who belonged to the local elite if there is one and how they "exploited" the court system to their benefit.
- Relations between the officials. For example, what is the relationship between the minister (*kirkkoherra*) and the sheriff (*nimismies*). Usually it's thought that officials were on good terms with each other, but my experience is that there was a lot of tension and disputes between them.
- Relationships at the micro-level (person-to-person). Looking at the networks from above is just one part of my PhD thesis. I will also study other aspects that reflect the state-building process. Even though I'm doing a lot of quantitative parts, I will also give room for qualitative examples. If I'm able to see, to which persons some individual was connected it will help me a lot to build the context around these qualitative examples.

Complex networks are hard for humans to grasp. However, since networks can be found in all fields of science (e.g. biology, economics, political science), they have been studied a lot and thus researchers have developed different tools and ways to visualize and analyse them. For this project,

[1] The 17th century Iisalmi included nowadays Iisalmi, Lapinlahti, Vieremä, Sonkajärvi, Kiuruvesi, Pielavesi.

I’m using Social Network Analysis (SNA) as a method to study networks. SNA is based on graph theory (mathematics) and there are several ready-to-use packages for it (e.g. iGraph, NetworkX). For my project, I chose an open-source software called Gephi to analyse and visualize the networks at the District Court of Iisalmi. Compared to other popular options it is fairly easy to use (because of the user interface) and the focus in Gephi is more on the visualization and less on the statistics, which was a good trade-off for my project (since I’m not going very deep level in the whole SNA).

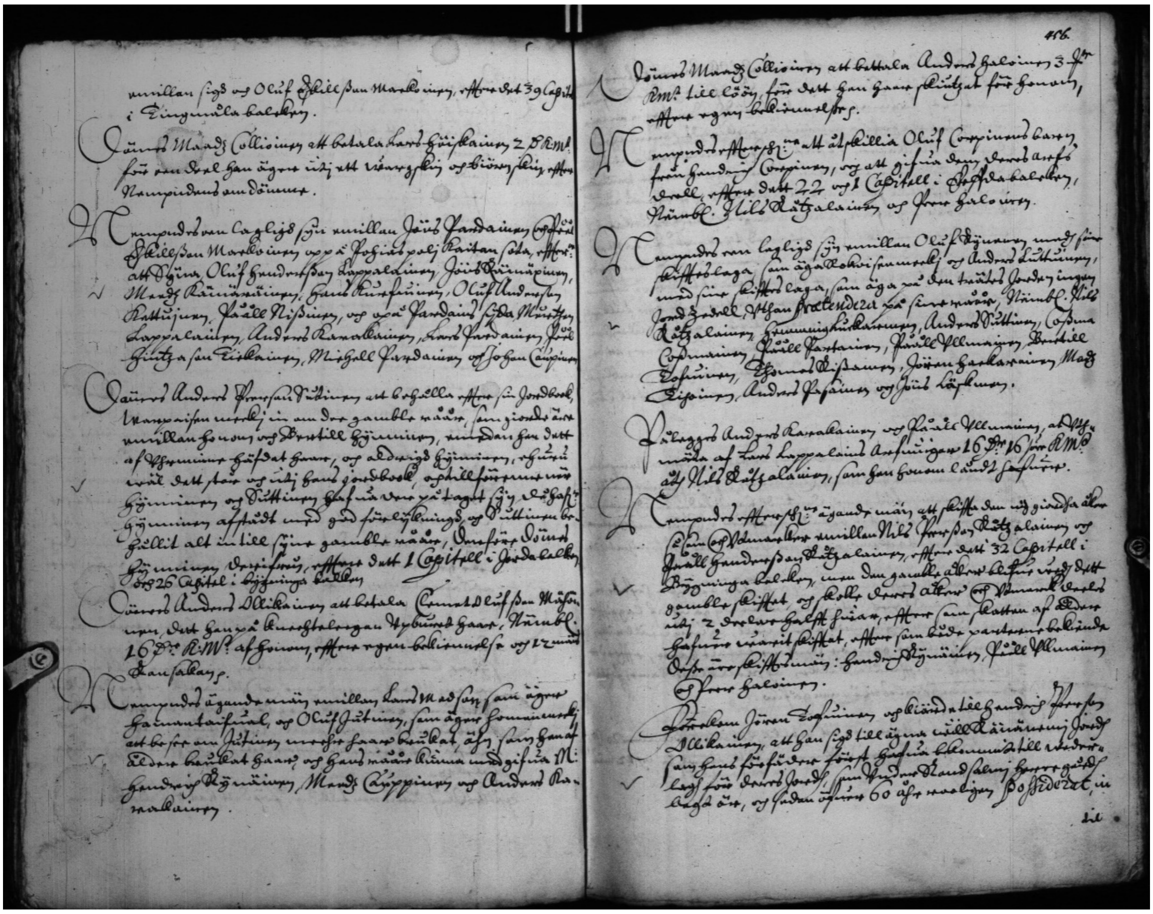
There is nothing new in applying SNA to historical data and it has become a popular method in the past few years. Many have already questioned this trend and point out that researchers who use SNA with historical data should also pay attention to source criticism, biases in the data, and evaluation (e.g. are the relationships between persons in the network real or false). It’s also often pointed out that humanists lack understanding in the theoretical background of SNA (graph theory), and thus blindly rely on the calculus made by mathematicians without understanding the principles and flaws that they come with (this is unfortunately also the case in this project).

My personal observation has been that usually the studies that involve historical data and SNA are looking at networks, which either have been already researched a lot or consist of “famous” individuals. For those type of studies, the results have been for most of the part very predictable (of course duplicating the previous results with new methods is an accomplishment in itself). However, in my research all the persons in my data are unknown as historical characters, so it’s not obvious who are for example the important hubs in the network. To be fair it’s expected that the “important” characters in the network are officials and the lay members of the court.

Data

The District Court Records of the Swedish Empire are unique and one of a kind source material. In the area of Finland continuous series begin from the early 17th century and continue to the late 20th century. Court Records are very diverse source material and one shouldn’t think that they are only for those who are interested in crime, justice or historical events. For example, it’s the best and when it comes to earlier centuries the only source material available if you are interested in such themes as everyday history, gender, sexuality and minorities. One of the reasons why Court Records are versatile as source material is that until the 19th century the district court (*käräjäoikeus*) wasn’t only a place where court and justice was held, but it also served as a local administration institution where local issues were discussed. That’s why most of the families (at least the head of the family) were present during the Court Sessions even though they didn't appear for any case. This collective and public nature of the district court is reflected in Court Records. Even though the justice system was very similar to modern times, the language and contents of the Court Records are very different from nowadays. Especially the earlier Court Records are by nature very gossipy and stray often to describe events that are not so related to the case. That's why Court Records are a treasury for historians and you never know what you can find from them. It's no wonder that historians have used them a lot and keep using them also in the future.

Picture 1. Example page of the Court Records (Winter Court Session, Iisalmi 1646).

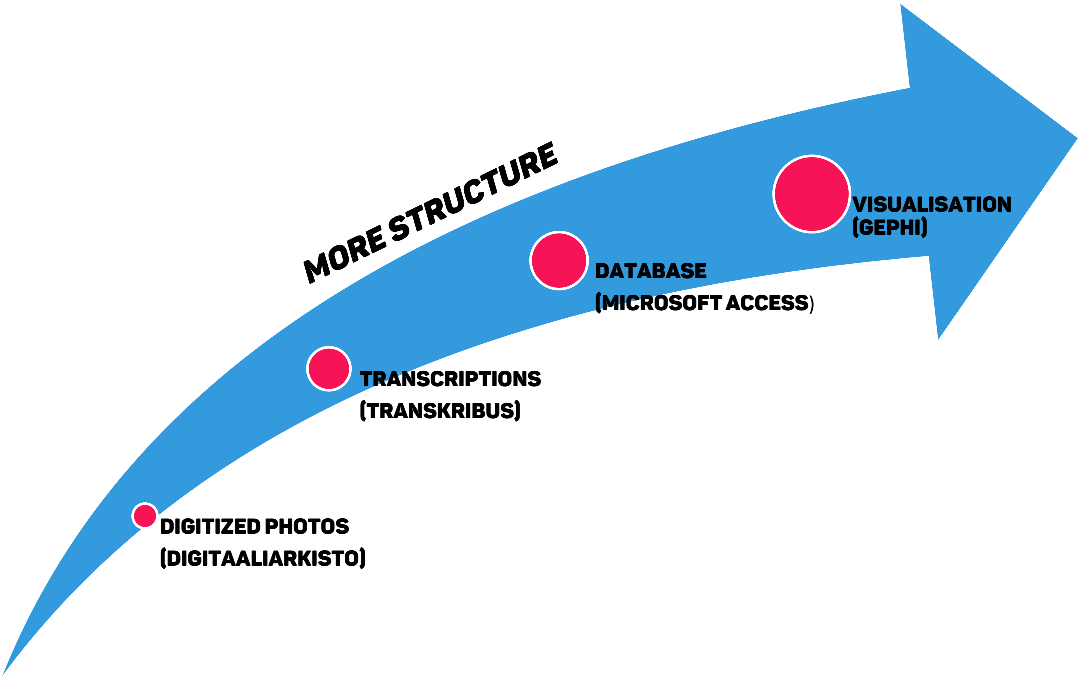


Data processing

Building the database that is used to make SNA (Social Network Analysis) has started long before this project. My plan has been to build a database that serves the whole PhD thesis and things have also changed a lot along the way (one might say there has been a "digital turn"). Some of the steps described here are not necessary if SNA is the only thing you want to do with the data. The data processing I have done can be divided into four steps which are portrayed in the Figure 1 below.

The goal has been that with every step the data is in a more structured form and thus it's easier to analyse regarding the research questions of my PhD thesis.

Figure 1. Data processing divided into four steps.



The 17th century Court Records have already been digitized and I have downloaded the ones I need from Digital Archives (*Digitaaliarkisto*) of the National Archives of Finland. Unfortunately, the digitization is done cheaply, and they haven't digitized the original Court Records but instead bad quality microfilms. However, since the resolution of the pictures is high some of the problems (e.g. very dark pictures) can be fixed with regular photo editing (adjusting local contrast, white/black etc.).

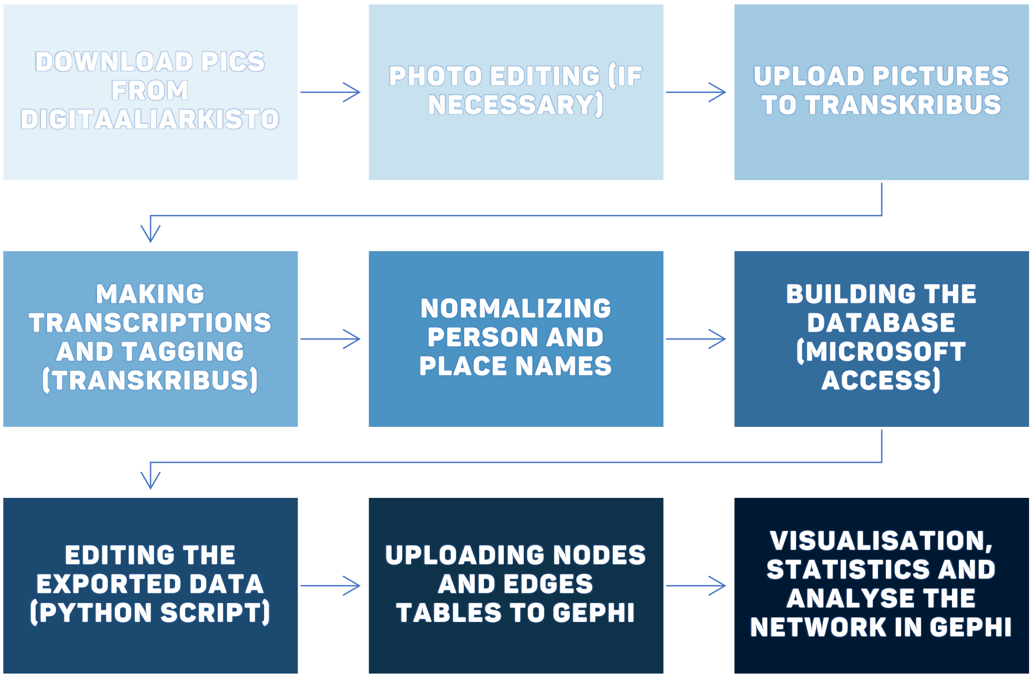
Doing the HTR (Handwritten Text Recognition) processing with a program called Transkribus was the first step to "enrich" my data by making the Court Records machine-readable and in a digital form. That is not necessary for the SNA but has helped a lot when it comes to building the database. If you are interested in the details of how the Court Records were transformed from analogical to digital with Transkribus see Appendix I.

After finishing the transcriptions, I started making a database with Microsoft Access, which increases the structure of the data by a lot. I didn't have previous experience with Access, and surprisingly there were a lot of things you had to learn the hard way. The major problem was that I lacked basic knowledge about building databases. I couldn't have done the database with Excel because I wanted a unique ID for every person, place and court case and also at the same time keep records about lots of different stuff. In the Access database I categorize the data case by case and take down following things:

- Metadata (year, court session, case number, page number)
- Main category (e.g. violence) and secondary category (e.g. assault)
- Short description of what was the case about
- Parties and other people mentioned in the case (plaintiff, claimant, defendant, witness, prosecutor, lay members assigned to the case, miscellaneous people mentioned in the case)
- For certain types of cases (e.g. right to land ownership cases) I categorise also extra information (e.g. name of the land, type of the land, type of the case, fines, compensation)
- In the database every case, person and place have a unique ID number.

For this project, I have narrowed the data to years 1639-1651 because so far, I have done my Microsoft Access database only for those years. However, it's fairly easy to reproduce the steps also for the whole database when it is finished.

Figure 2. Data processing pipeline.


DOWNLOAD PICS FROM DIGITAALIARKISTO
PHOTO EDITING (IF NECESSARY)
UPLOAD PICTURES TO TRANSKRIBUS
MAKING TRANSCRIPTIONS AND TAGGING (TRANSKRIBUS)
NORMALIZING PERSON AND PLACE NAMES
BUILDING THE DATABASE (MICROSOFT ACCESS)
EDITING THE EXPORTED DATA (PYTHON SCRIPT)
UPLOADING NODES AND EDGES TABLES TO GEPHI
VISUALISATION, STATISTICS AND ANALYSE THE NETWORK IN GEPHI

After exporting the data from Access, the next step was to re-arrange the data since in the database it is stored case by case and SNA requires that the relationships (edge table) are between persons (node to node). This task wasn’t super hard to solve. I think it is also possible to do with Excel somehow, but I ended up doing it with a self-made Python script. The code wasn’t super complicated, and I needed only the CSV (Comma Separated Value) library outside the Python Standard Library. Most of the code could be done with the basics of Python and I copy-pasted code (from GeeksforGeeks) only in the part where I needed to make all possible (and unique) pairs out of persons in a list.

In order to do Social Network Analysis, you must determine what are the nodes (actors) and edges (relational ties) in your network. For my project, I consider that the nodes are the persons in the court cases. For the edges, I consider that all persons inside one court case have undirected “relationship” with each other and thus they have an edge between each other.

It’s logical for example to think, that the parties (plaintiff and defendant) have some sort of relationship that has led to a situation where they are in the courtroom facing each other. However, this is a huge simplification and doesn’t come without problems. For example, there are a lot of cases that are about land-owning rights and lay members of the court (*lautamiehet*) are sent to solve these disputes. Can you consider that there is a real relationship between the lay member and the parties? Then there are the witnesses for example. Sometimes it’s clear that they have connections to both parties but sometimes the case might be that they don’t really know each other.

I have also listed a lot of people under miscellaneous roles. Sometimes they are mentioned because of a small detail (e.g. they have been the previous owner of some land lot) and sometimes they are very close persons in that case but can’t be categorised to other roles (e.g. the previous partners in adultery cases). For this project, I didn’t do any weighting for the edges, so at the moment the relationship between plaintiff and defendant (which can be considered to be a strong relationship) is weighted as much as the relationship between witnesses who might not even know each other. The reason why I didn’t do the weighting is that for me it feels quite arbitrary; what values should one use? So, in order to do that I need to find some examples.

Analysing the data in Gephi

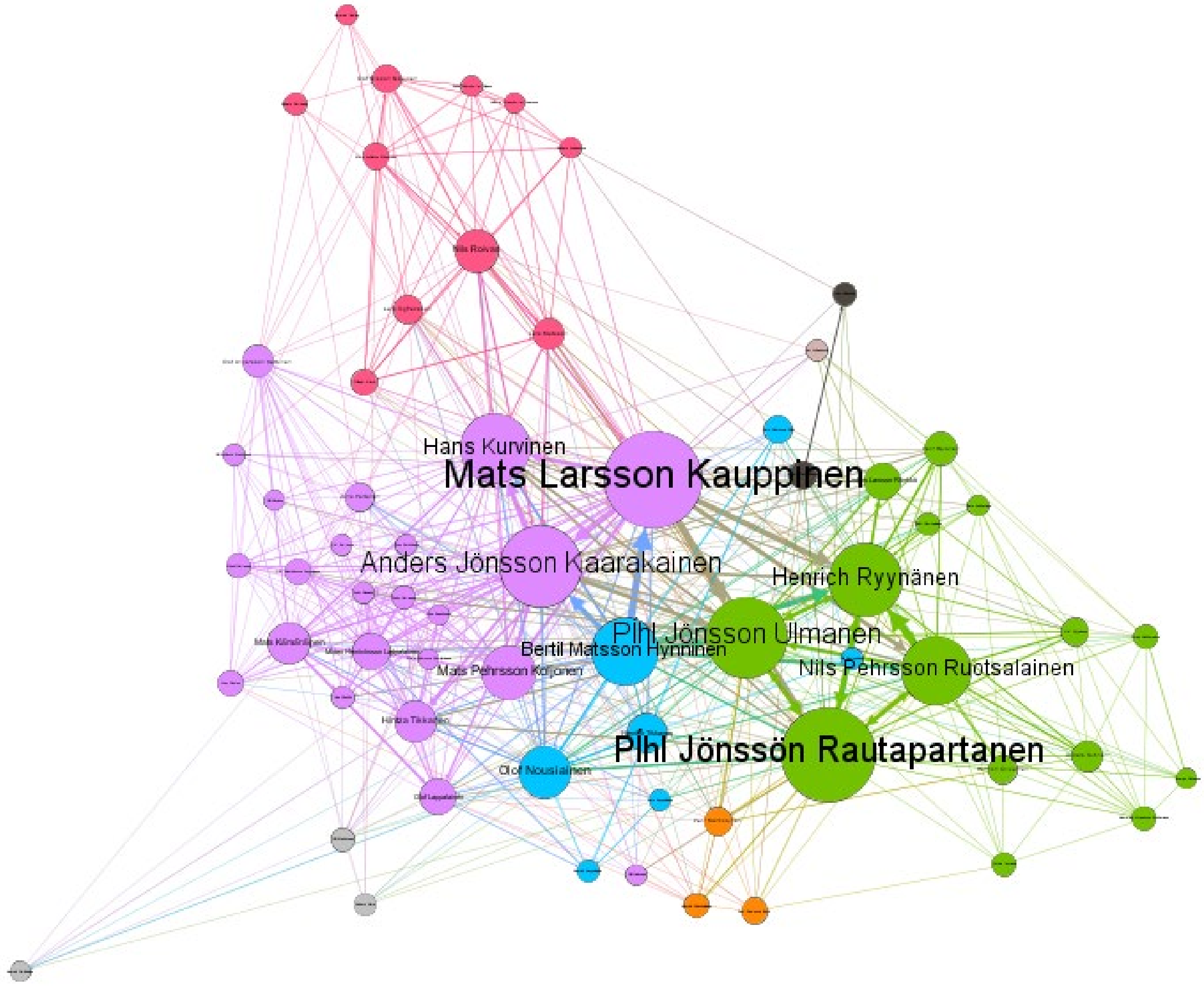
I would have definitely needed more time to do properly the analysing part of the project (which is of course also the most important part). The problem was that SNA (Social Network Analysis) is a relatively new method to me and I haven’t done much reading about the statistical and theoretical part of it (graph theory). So, like many others before, I made the mistake that I rely on algorithms and theories whose background and meaning I don't truly understand.

I think Gephi was the right choice for my project since I managed to understand the basics only by watching a few short online tutorials on YouTube. Importing the node and edge table was straightforward even though for the first time I made a serious mistake and thus every edge between the nodes was listed as directed relations even though the relations in the Court Records (for most of the time) are undirected.

A lot of time in Gephi was used to play around with the network. My presumption was that since SNA is mostly developed by those who come from "hard sciences", the results should be universal and reproducible. Of course, I knew that a lot of things can be emphasized with colour, size and perspective. However, what I quickly realized about SNA is that there are countless amount of options, and altering any of them will strongly affect the outcome. For example, there are multiple options to do the layout (how nodes and edges are positioned) and the most popular ones (Yifan Fu and Force Atlas) have also many parameters that can be altered to get different results. You can also display the network with unlimited different ways (node colour, size, label). Also changing most of these parameters also greatly affects the statistics, which can also be done in the program. Of course, when you run the statistics you can also change parameters. So, my first impression was that SNA is a quite arbitrary method and the idea is more that you get to choose how you want to represent the data. So as a historian who often hears that history is just storytelling (which I totally agree on) these methods that have a background in "hard sciences" seemed more random. Of course, when you think about it's quite obvious that there is not a universal way to study networks and it´s expected that results vary a lot depending on how you approach it and what the research questions are.

One of the first Networks I managed to do is shown below (Network). The first outcome was predictable, and I saw it coming. The network is built around the lay members of the court (*lautamiehet*). There are several reasons for this, but the main reason is that the lay members of the court were sent as receivers (*selvitysmies*) to solve disputes. During that era, there were 12 lay members in the court at a time and for my time period 1639-1651 there are around 15 of them. However, receivers were not always lay members of the court and often also regular peasants were used as receivers. In my edge table, I have recorded that for every case that includes receivers, all of them have a relationship with each other and when this happens many times (amount of receivers sent to solve one issue varied from 2 to 12) the relationships between them are highlighted. Despite the obvious problems the first network already told a lot of information about the local community.

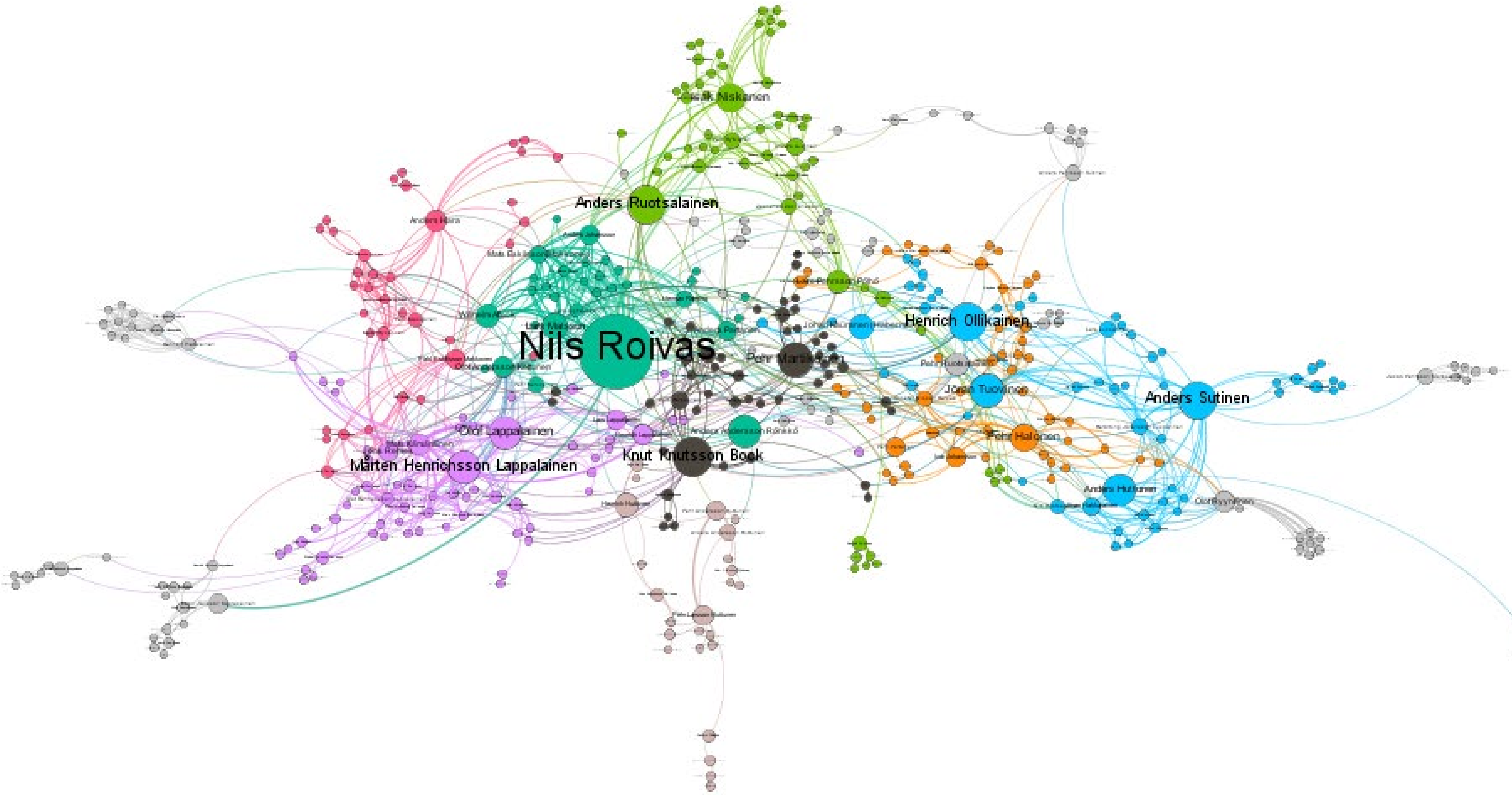
Network 1. Iisalmi District Court 1639–1651. Filtered with degree range (only part of the network is shown). Different clusters (colour) are grouped by modularity.



However, I already knew that I would run into trouble. One of my main aims in the first place was that the network should be filterable by:

- Status (e.g. exclude all the gentry out, and leave only peasants)
- Roles (e.g. see only relationships between plaintiff and defendant)
- Category and sub-category of the case (e.g. see the network if we only look at land ownership disputes)

Network 2. Iisalmi District Court 1639–1651. Lay members of the court excluded. Different clusters are grouped (colour) by modularity.



Filtering the network is pretty straightforward in Gephi. The most useful filter is Degree Range (how many connections one node has), which can be used to filter not so significant nodes from the network (and make it look less like a hairball). You can also run the statistics so that you only take into consideration the filtered data.

There is also an option to filter the network by attributes in the node and edges table. This is great because I have for every edge information about the case category and type of relationship. For some of the nodes I have also the status (e.g. bailiff).

Above (Network 2) you can see the result if we exclude completely the lay members of the court. It's no surprise that after this the most central person appears to be the minister Nils Roivas. But now you can also see other peasants that are central to the network (e.g. Henrich Ollikainen). While you see a lot of new things from this perspective, I don't think that excluding the lay members of the court like this is good if you want to examine the community as a whole. Because the lay members of the court were peasants, they appeared in many cases as regular people so excluding them entirely will affect the network significantly.

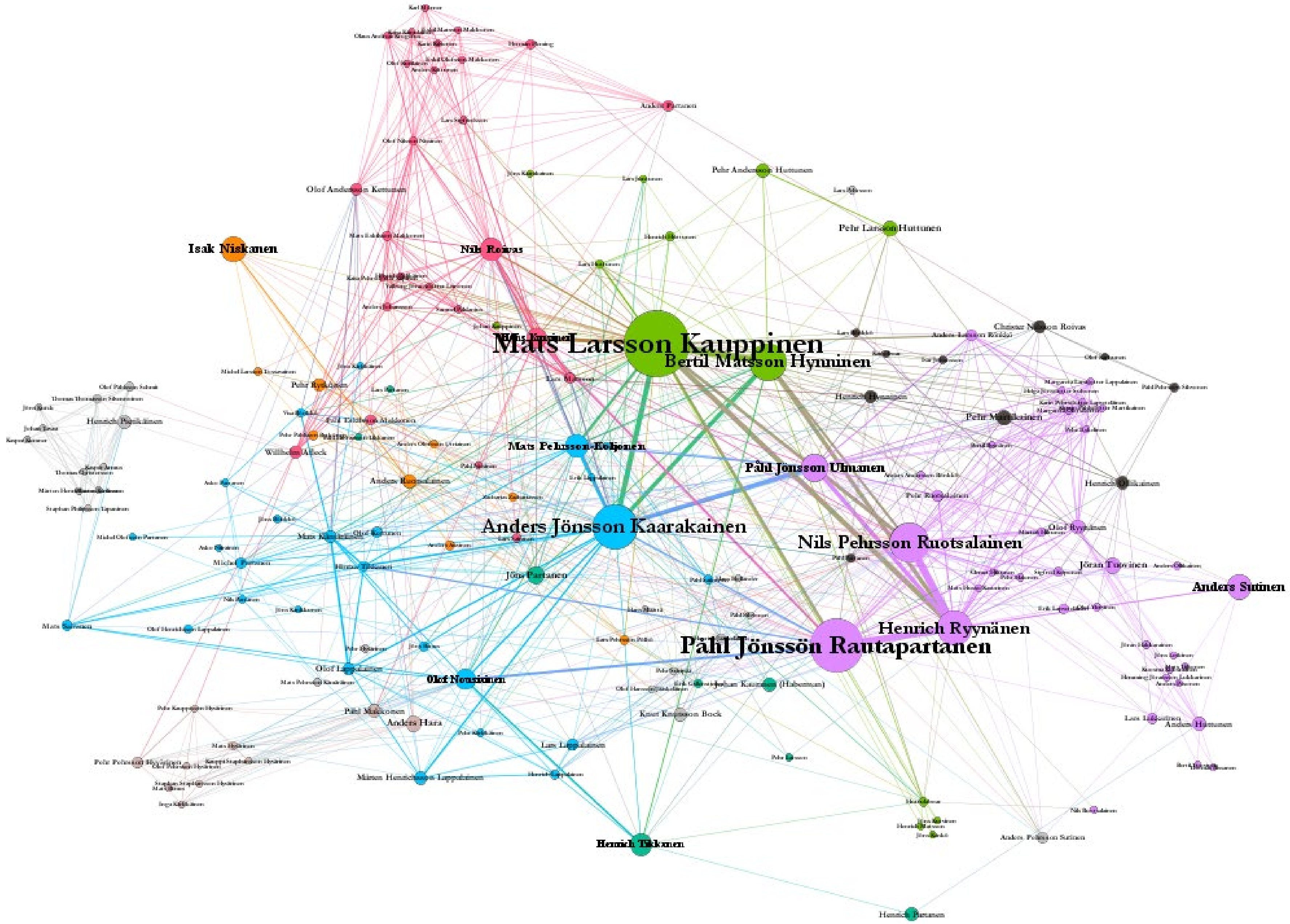
That's why the next step was to filter the network by the type of relationship. Below (Network 3) you can see the network if we exclude the relationships between receivers (*selvitysmies-selvitysmies*). The lay members of the court are still central to the network, but they are not anymore equally significant and there are also other peasants in the network who have high centrality betweenness.

Here are some statistics about the network:

Statistic	Value	Term explained[2]
Average Degree	9,086	Average number of links per node
Average Weighted Degree	12,613	Average of sum of weights of the edges of nodes
Network Diameter	8	The maximum distance between any pair of nodes in the graph
Average Path length:	3,119	The average path length is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of the efficiency of information or mass transport on a network.
Average Clustering Coefficient	0,760	A measure of the degree to which nodes in a graph tend to cluster together

[2] Social Network Analysis, https://sites.google.com/a/umn.edu/social-network-analysis/

Network 3. Iisalmi District Court 1639–1651. Receiver- receiver (*selvitysmies-selvitysmies*) edges excluded. Different clusters are grouped (colour) by modularity.



Potential bias and problems

I have already pointed out that there are many problems when it comes to the edges (relationships). Doing them better would take a lot of time and whilst doing research you have to always balance between efficiency and accuracy.

There are several biases considering the Court Records:

- Some of the Court Records are destroyed, some included in the dataset have bad quality. Thus, my database doesn't represent all the Court Records of Iisalmi.
- It depended a lot on the scribe and time what is written in the records (e.g. late 17th century Court Records are much longer and include more information compared to ones written in the early 17th century). In my opinion, it is very arbitrary what persons are mentioned in the Court Records. Scribe can be explicit and tell who for example gave testimony about one's reputation. On the other hand, it's common to use the passive form and just say that the reputation was proven by some individuals in the courtroom (*käräjärahvas*).

There are also several biases considering my database:

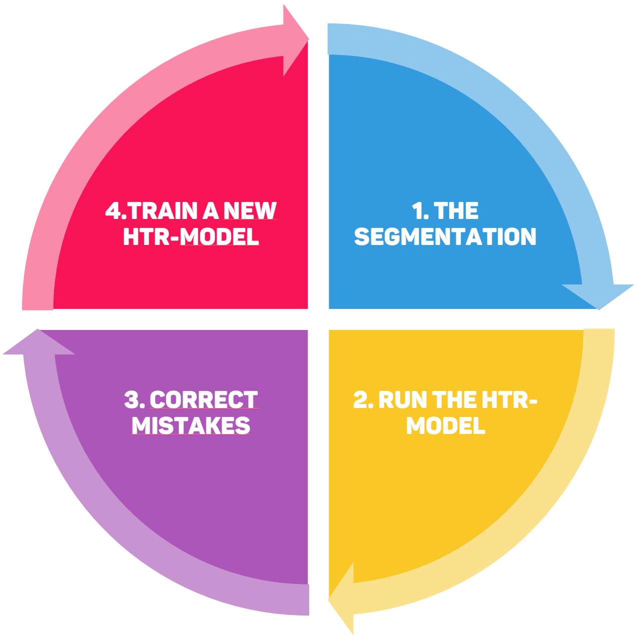
- Categorizing the status of persons inside one court case has several flaws (e.g. some persons could be categorized into many categories). It should also be noted that the legal terms (e.g. plaintiff) are rarely used so it's mainly my own interpretation as a historian who doesn't have any studies in the faculty of law. There is always also the question should one even try to interpret historical data with modern terminology (anachronism).
- Identifying all unique persons is impossible. For example, it's hard to tell if Mats Matsson Lappalainen and Mats Lappalainen are the same or different person (the fun fact is that usually the same name is given to the first son). However, Northern Savonia is perhaps the only place where you can even try to do what I'm doing because they already used surnames for peasants (most of Finland it is just a first name and patronymic name) and the size of the community is manageable. Also, in many court cases the context helps a lot and thus I can tell that who are the unique persons.
- It wasn't uncommon that already deceased persons are mentioned in the court records (e.g. some peasant's ancestor who has reclaimed some land lot). At the moment these people are listed the same way in the network.

- One bias is also that the chosen years (1639–1651) doesn’t actually represent anything. However, this is not a major problem and I can redo the analysis when I have the database for the whole 17th century.

Appendix I: HTR-process in Transkribus

Since I wanted to have 100 % correct transcriptions the reason to use Transkribus and HTR was to speed up the process. That is why the work done in Transkribus was a cyclical process and can be divided into four parts (see Figure 3 below).

Figure 3. The HTR processing in Transkribus when you want 100% correct Transcriptions.



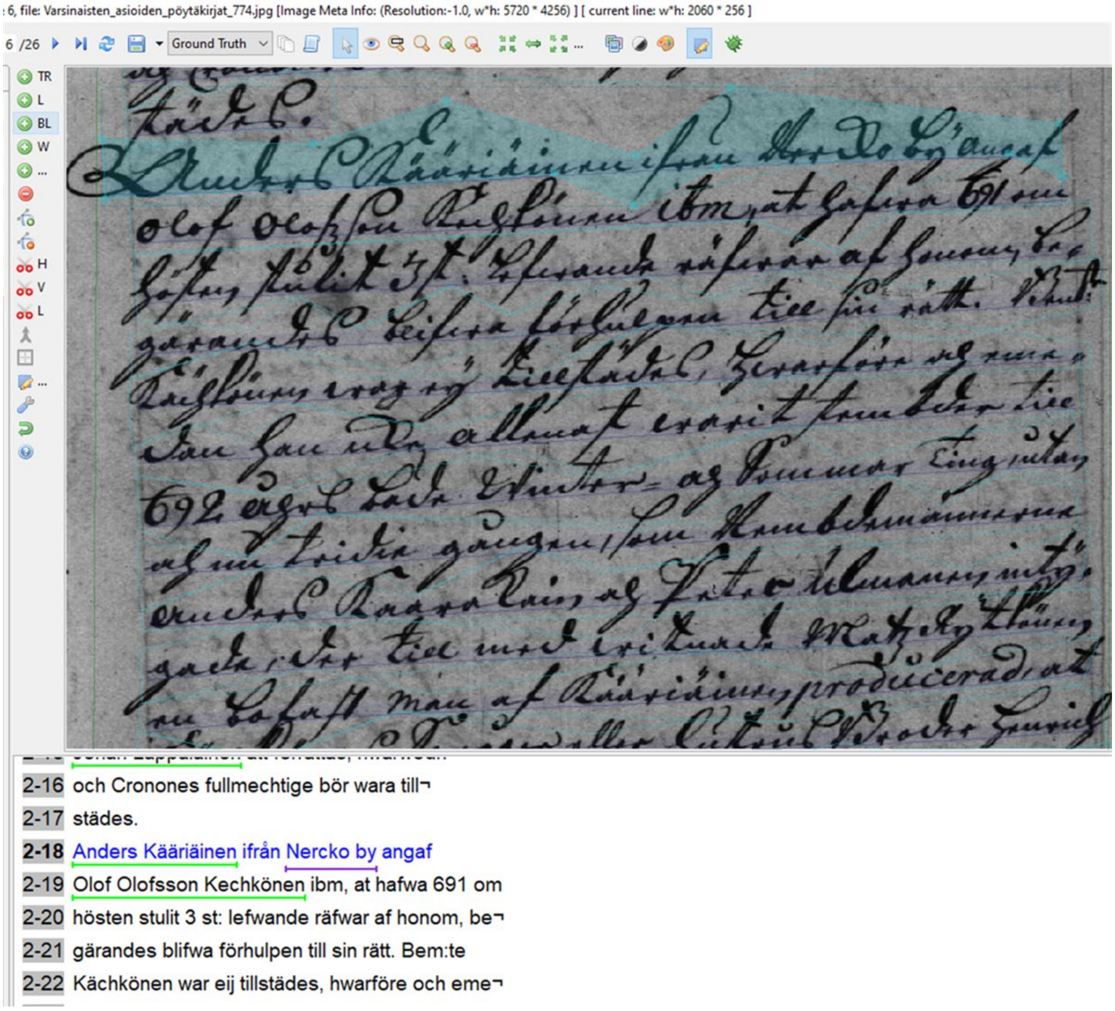
Before the first part you upload the pictures to Transkribus (and if necessary, you must do photo editing before this). The first part is that you do the segmentation (i.e. where the lines of text are in the picture) by applying the automatic Layout Analysis and correcting the mistakes that it might do. After that you apply your latest HTR-model (if you don't have a model yet you fill the empty lines). Then you correct the mistakes made by automatic text recognition and thus it becomes a new training set (Ground Truth) for the next model. When you have done enough Ground Truth you make a new HTR-model.[3] It's expected that the new model will have a better accuracy and thus in theory this cyclical process becomes faster and faster with every cycle because there are fewer mistakes to correct. In practice, the accuracy of the HTR-model doesn't always get better if you have data, which has several different scribes. Usually the HTR-model struggles with new handwriting,

[3] In my case I made a new HTR-model usually after every Court Session, so around 10-20 pages. Actually, if you want to be most efficient you shouldn't do this in a linear fashion (i.e. go one document at a time) but choose the Training Set representatively (e.g. one page from each document). However, I did this "wrong" way because I valued more that you get the overall picture whilst doing the transcriptions.

but after making some Ground Truth for that hand it gets better. However, when you have a lot of Ground Truth your HTR-models' ability to read hand-writings that are new to it increases, and at some point it should also be able to read those as well as ones that are already in the training set.

Whilst correcting the mistakes made by automatic text recognition, I have also tagged three things. Firstly, I have tagged every first letter of every court case. By doing this I can quite easily separate court cases from each other (there isn't any regular indicator in the text that could be easily used as a separator).

Picture 2. Screen-capture from the Transkribus platform. The tagged persons have green underline and tagged places have a purple underline.



: 6, file: Varsinaisten_asioiden_pöytäkirjat_774.jpg [Image Meta Info: (Resolution:-1.0, w*h: 5720 * 4256)] [current line: w*h: 2060 * 256]

2-16 och Cronones fullmechtige bör wara till¬
2-17 städes.
2-18 Anders Kääriäinen ifrån Nercko by angaf
2-19 Olof Olofsson Kechkönen ibm, at hafwa 691 om
2-20 hösten stulit 3 st: lefwande räfwar af honom, be¬
2-21 gärandes blifwa förhulpen till sin rätt. Bem:te
2-22 Kächkönen war eij tillstädes, hwarföre och eme¬

Secondly, I have tagged every person name. The challenge with these is that conventionally to the period there can be many different ways to spell the same name. Also, the format varies a lot and for example the same person can be Mats Andersson, Mats Partanen and Mats Andersson Partanen. I have solved the first problem by normalizing all the names. This was done by first splitting every tagged name by space and removing duplicates. After that, I assigned “normalized form” for every unique name. After this I (with help of a friend) a simple Python program was used to normalize the names.

Thirdly, I have tagged every place name. Place names are also normalized in the same way as person names. However, there still are certain challenges considering them. Since I have tagged every place name in the same way the output includes nation names (e.g. Tyskland = Germany), province names (e.g. Savolax = Savonia), parish names (e.g. Wiborg = Viipuri), village names and names of homesteads, plots, lakes, rivers, forests, fields and meadows. So, the data is very heterogeneous, and it would require a lot of work (and in many cases it would still be impossible) to for example get coordinates for every place mentioned so you could map the data. However, I’m interested to try “mapping” in the future, because I think it would open a whole lot of new possibilities (and the question of distance is an important factor whilst studying peripheries).

Some information about the data:

- 388,000 total words and 41,000 unique words
- 94 things (käräjät) and around 2,000 court cases
- Around 3,300 unique persons and 1,700 place names