Journal of Genetic Genealogy

Journal: www.jogg.info
Originally Published: Volume 9, Number 1 (Fall 2021)
Reference Number: 91.006

AUTOSOMAL MATCH IN-COMMON-WITH CLUSTER ANALYSIS WITH NETWORK VISUALIZATION TOOLS: AN EXAMPLE USING THE GEPHI OPEN-SOURCE TOOL

Author(s): J. David Vance

Journal of Genetic Genealogy

AUTOSOMAL MATCH IN-COMMON-WITH CLUSTER ANALYSIS WITH NETWORK VISUALIZATION TOOLS: AN EXAMPLE USING THE GEPHI OPEN-SOURCE TOOL

By J. David Vance

Abstract

The genetic genealogy community has many tools for autosomal DNA analysis, and many tools and techniques have been developed to use autosomal DNA match inter-relationships to assist in the identification of common ancestors. Many of these techniques work best with matches who share larger amounts of DNA and are therefore closer relatives whose genealogical connections are more readily discovered. This review discusses the merits of yet another technique, network visualization, which can cluster large datasets of matches even lower than 20 cMs (in this review, down to 7 cMs) and can identify and analyze clusters of In-Common-With matches which, especially when combined with other genealogical information like known relationships of certain matches or clusters determined by other methods, can help focus and prioritize our analysis of matches to find our shared ancestry and thereby extend our genealogical knowledge. In this review the Gephi tool was used as the network visualization platform but the approach is independent of the specific tool.

Background Assumptions

This review is not intended for those new to autosomal DNA analysis; not because the techniques are difficult to understand but because there are more commonly-suggested starting analysis techniques like the Leeds Method or even the analysis tools provided by commercial companies, and this review covers an approach which might be more useful after those starting techniques have been exhausted.

For that reason, this review does assume that readers have a basic familiarity with other autosomal DNA match analysis techniques like the Leeds Method and some fundamentals of autosomal DNA analysis for genealogy like the relationship of shared autosomal DNA segments to genealogical relationships between matches.

Introduction

When many of us take our first autosomal DNA (atDNA) test what we are hoping to find are matches who will help us figure out the gaps in our knowledge of our own ancestry. We hope for as large a pool of matches as possible with the somewhat mixed blessing that we then have to untangle the often difficult questions of how they all may be related to us and to each other.

Very often a key subset of matches will share larger centimorgans (cMs) of DNA with us and our relationship with those matches will be closer and clearer (say, perhaps within 3rd cousins). For this subset, where the genealogical relationships between ourselves and those matches don't

become clear by simply comparing our known genealogies we can often ferret out the relationships using our better-known atDNA analysis techniques: Leeds Method, segment matching, and so on, or using the suite of deservedly-popular tools which have been developed by the commercial companies and third parties in support of those techniques.

The success rate of our most common techniques though drops off rapidly with more distant relationships and as the shared DNA segments get smaller. While sometimes there is no substitute for doggedly researching and comparing genealogies to find common ancestors, these more distant matches can still be frustrating for genetic genealogists especially if there are a large number of matches in the "4th cousin and further" category whose genealogical relationships are unclear and who are especially resistant to analysis with our most common techniques.

A lesser-known approach especially for tackling these more distant matches is analysis using network visualization software to group them into In-Common-With (ICW) clusters – groups of matches who are themselves matches to each other and who may as a result all descend from a common ancestor.

This is not a new approach – network visualization approaches have been used for ICW cluster analysis at least as far back as 2017 by Barbara Griffiths and Shelley Crawford using a variety of tools including Pajek and NodeXL. In this review we have used the Gephi tool (free and open-source at https://gephi.org/) as the clustering and graphing platform but except for different flexibility in clustering and filtering options, the approach is independent of the tool and any similar network visualization software package would support the same approach.

Others are also applying network visualization to their own autosomal data analysis using similar but not identical approaches to the examples shown in this review. Their results are often displayed on social media forums to other genetic genealogists and generate much surprise and discussion, which suggests that the techniques are not widely practiced and might benefit more people if they were more widely understood.

Network visualization is not a replacement for more common analysis techniques; in fact as a clustering approach for ICW matches it is very similar to a Leeds Method analysis though more complicated to set up the necessary data and to analyze. This is one reason that a simpler method like Leeds should be attempted first, but another reason is that as we will explain in this review, mapping an initial subgroup of matches to their genealogical relationships using other methods first can be extended by network visualization to wider clusters of more distant matches. This means that network visualization is not only a stand-alone technique but also an approach that can extend the results of prior analysis.

The other advantage of network visualization is that it can be used to very quickly sort large networks into clusters and then explore these clusters to investigate their shared origins, and to analyze the network as a whole through the application of different filters that highlight important relationships both within and across clusters.

While we present a few examples of network visualization analysis of ICW clusters in this review, we would also propose that network visualization should be considered a more general approach which may include several analysis techniques depending on desired outcome and number and type of autosomal matches. Our main point in

writing this review is simply to show by example the merits of network visualization as an important tool in an analysis toolkit, not to suggest that there is one best network visualization analysis approach.

In this review we show by example how a network visualization tool like Gephi can be used to sort a large ICW network into clusters which include matches at smaller levels of shared cMs. We offer one approach for extending the relationship knowledge of a small number of matches out to the rest of the identified clusters, and also show how the tool's filtering options can be used to dissect the network in different ways to gain additional insights.

Methods and Data

Preparing the Network Data

A network for Gephi purposes is simply a set of "nodes", each connected to other nodes via "edges" which are represented by connecting lines. For the tool to represent a network the minimum required data therefore are a "node table" which lists all the network's nodes, and an "edge table" which has pairs of node ids representing the two endpoints of each line in the network.

At a minimum then this could be represented by two tables, a Node Table as in the example in Figure 1, and an Edge Table as in the example in Figure 2.

Id
N00001
N00002
N00003
N00004
N00005

Figure 1a. Simple Node Table

Source	Target
N00001	N00003
N00001	N00005
N00002	N00003
N00002	N00004
N00003	N00002
N00003	N00004
N00004	N00002

Figure 2. Simple Edge Table

In this example we will use an undirected network, so which node is "Source" versus "Target" is immaterial (and note that while there are duplicate connections shown in the edge table in Figure 2, Gephi removes those on undirected networks). More complex network analyses however could be conducted using directed networks or even weighted edges (for example weighted by number or size of shared cM segments, etc).

To build this network we will use the match lists provided by an autosomal DNA testing company. Most of the major testing companies (Ancestry, Family Tree DNA, and 23andMe as examples) provide this level of detail, although not all of them provide it in easily downloadable files.

The Node Table of course is simply our list of autosomal matches, while the Edge Table lists which of our matches also match each other. These lists can be built by hand, though the commercial companies do not all provide easy identification of matches that also match each other.

A more automated method of producing these files is to use the DNAGedcom tool; their "match" and "icw" (in common with) output files can be used for Node and Edge tables, respectively, for this purpose and only columns headers need to be changed (Gephi requires "Id" as the header in the Node Table

and "Source" and "Target" as the headers in the Edge Table). At the time of this publication not all commercial companies however allow the use of DNAGedcom to download their data, so this should be investigated before attempting that method. Or ICW information can also be obtained from an automated cluster assessment from DNA Painter or Genetic Affairs.

Brit Nicholson has an excellent tutorial on building these files from GedMatch data in his Aug 2020 blog post on https://www.dna-sci.com.

Otherwise these files can also be built manually using the reports on the testing companies' websites (For example the Shared Matches report for AncestryDNA, or the Relatives in Common report from 23andMe).

For the example in this review, DNAGedcom was used to create the original ICW data files.

One powerful addition to the network data is the ability to include additional columns in the Node Table and use them for additional analysis. For our actual example in this review, we will add four additional columns:

1. Match names, which normally would be the given names of matches but will here be represented by the labels "Match #1", "Match #2", etc.;
2. Known Genealogical Relationships, discussed in the next section;
3. Shared cM between that match and our DNA, so that the network can include this for filtering and analysis purposes;
4. The cluster numbers resulting from a Collins-Leeds Method analysis of the matches with larger shared cMs; this was done only for test purposes to show how network visualization compares to Leeds Method approaches and is not otherwise necessary for network analysis.

These additional columns are shown in Figure 3.

Id	Name	Known Relationships	Shared cM	CLM Cluster
N00001	Match #1	1-1-2-1	263.4	25
N00002	Match #2	1-2	892.6	25
N00003	Match #3		45.9	6
N00004	Match #4	2-1-2-1-2	123.7	18
N00005	Match #5		27.6	11

Figure 3. Adding More Columns to the Node Table.

The "CLM Cluster" column used in this example was added solely for this review and would not normally be required. The Collins-Leeds Method was run on the same match data for comparison to network visualization; as shown later in our example, clusters using Gephi match up with clusters identified through Leeds Method analysis but can be more easily filtered to visualize and highlight relationships.

To populate this column, the Collins-Leeds Method identified clusters for some 2,870 matches and these cluster numbers were listed for those matches in a new column of the Nodes Table as shown in Figure 3. This data will be used later in this review as a comparison between clustering methods.

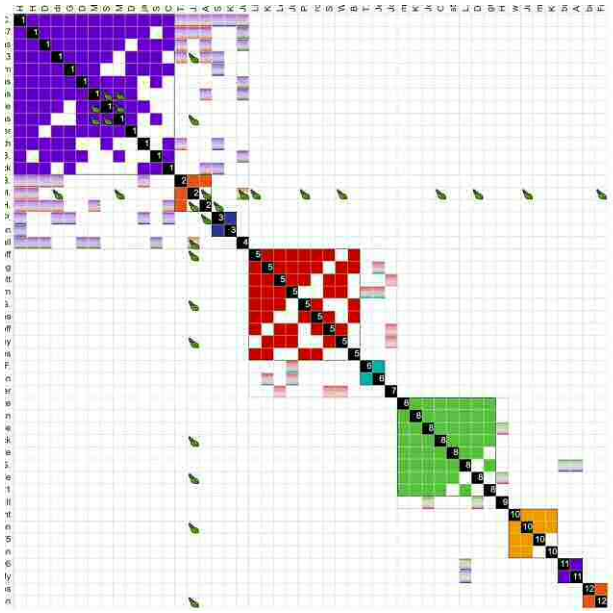

Figure 4. Collins-Leeds Method run on the ICW data for comparison purposes. This was used only to identify common clusters for a subset of ICW matches (names have been omitted for privacy).

Representing Known Genealogical Relationships in the Data

Usually with our match lists there is at least a small subset where we already know the genealogical relationships between us and those matches. That subset may include known relatives, matches with whom we have compared genealogies and identified the common ancestors, or even matches whose genealogy we have built out through traditional research methods and have been successful in identifying the connections.

If this subset exists, the known genealogical connections can be helpful in identifying the common ancestors of clusters and, by extension, the common ancestors of other matches to each other and with ourselves.

To demonstrate this in this example, we are using a modified Ahnentafel numbering system, where male ancestors are represented by "1" and female ancestors represented by "2", and each generation is represented. Therefore our father is "1", mother is "2", father's father is "1-1", father's mother is "1-2", and so on. Figure 5 also includes an example match on the left who descends from our great-grandparents on our maternal grandfather's side. Since both we and the match inherited DNA from the pair of ancestors marked "2-1-1" and "2-1-2", we represent this as "2-1-1/2" in the Known Relationships column.

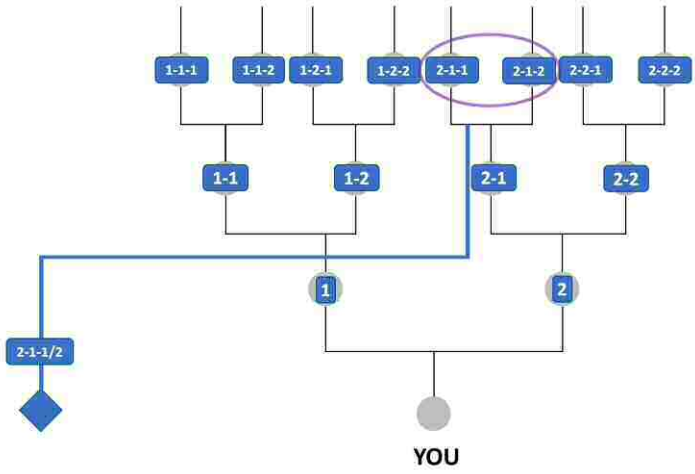

Figure 5. Keeping track of Known Relationships

The purpose of this column will become more clear in the example below, but this nomenclature was selected because it identifies the shared ancestral lines and number of generations between us and certain matches.

In our example, we have identified known genealogical relationships for approximately 200 of the closer matches and this information has been captured using this custom nomenclature.

Combining these methods for one of the author's autosomal DNA tests resulted in two files to use as input for our example: a Nodes file of 31,758

matches with extra columns as shown in Figure 3, with some 200 of these matches marked with Known Genealogical Relationships and 2,870 of the matches marked with a CLM Cluster number from the Collins-Leeds Method clustering mentioned earlier.

Creating the Network

Installation of the Gephi tool is beyond the scope of this review, but requires the correct Java runtime environment and enough CPU to handle the intensive computational requirements. This example was created using Gephi version 0.9.2 on a Microsoft Surface 3 running Windows 10 with an i7-1065G7 CPU and 3GB of RAM.

The tool allows the import of data as a Node Table followed by the Edge Table; and when the data is first imported it is displayed in the Graph display and initially appears as a square without identifiable structure or color, as shown in Figure 6.

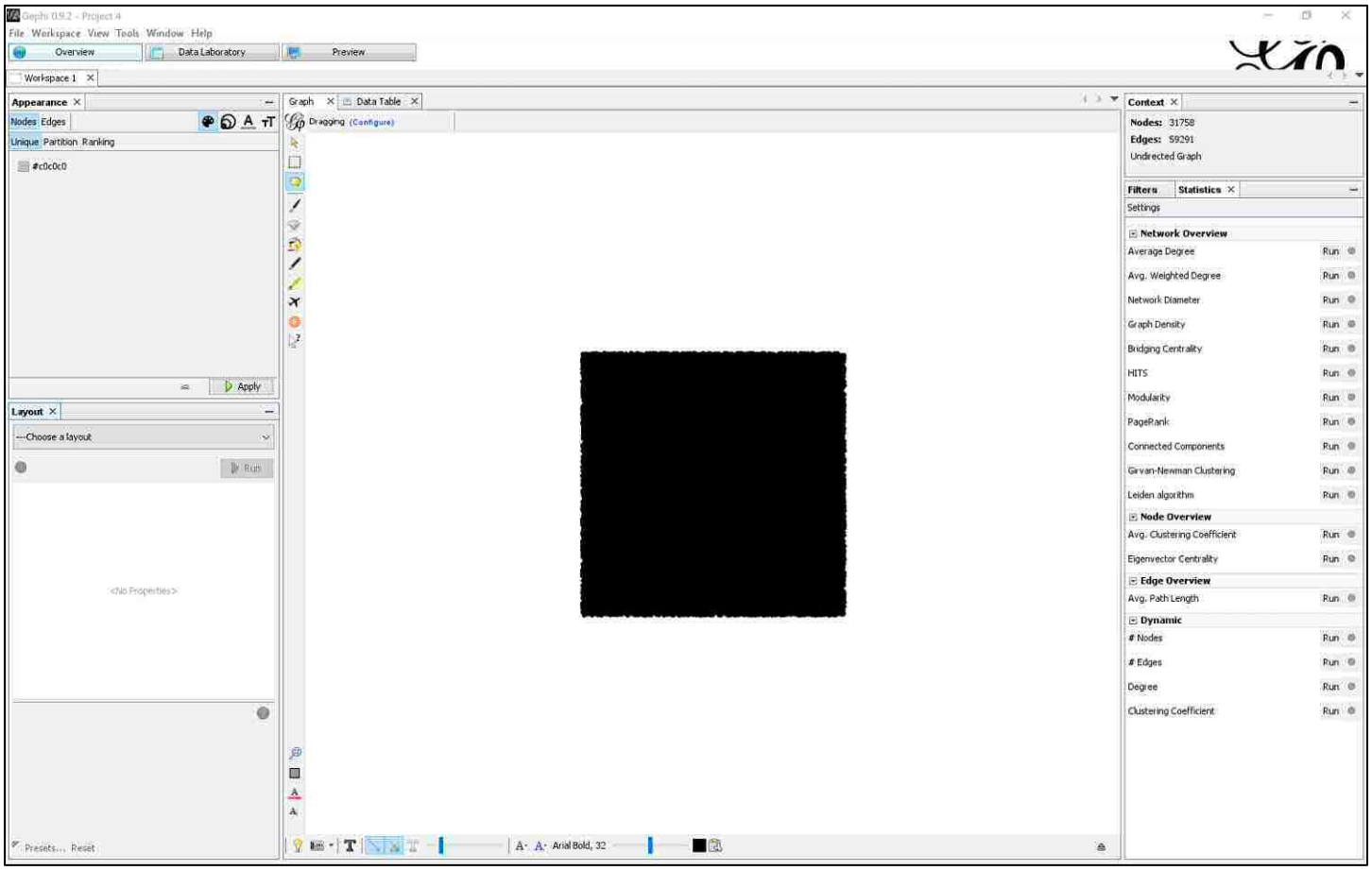

Figure 6. What a network in Gephi first looks like.

A complete overview of Gephi functions is also outside the scope of this review; in brief the "Appearance" window in the upper left controls the appearance of the graphical network display – colors, size of nodes, etc; while the "Layout" window in the lower left controls the clustering

analysis methods that can be applied to the network. In the upper right the "Context" menu gives information about the current network or portion thereof which is currently being displayed, while the "Statistics" menu in the lower right allows various statistics about the network to be calculated. The "Filters" tab to the left of the "Statistics" tab further allows the flexible application of a number of filters. These options will all be important in this example.

Applying a Layout

In the lower left window, Gephi will show various layouts that can be applied to the network to arrange the nodes in the 3D display graph according to various criteria. Which one to use depends on the characteristics of the network, though for networks of ICW matches the "Force Atlas 2" layout appears to work best. In simple terms, this spatialization algorithm causes nodes to "repulse" each other and higher numbers of edges between nodes to be "attracted" to each other and as the algorithm is run the network visualization stretches in real-time into clusters of nodes which share higher numbers of edges with each other. This algorithm was developed by the authors of Gephi and is more fully described in this published study: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679.

After applying the Force Atlas 2 layout (and after zooming the graph display out to include all nodes) the graph looks like Figure 7.

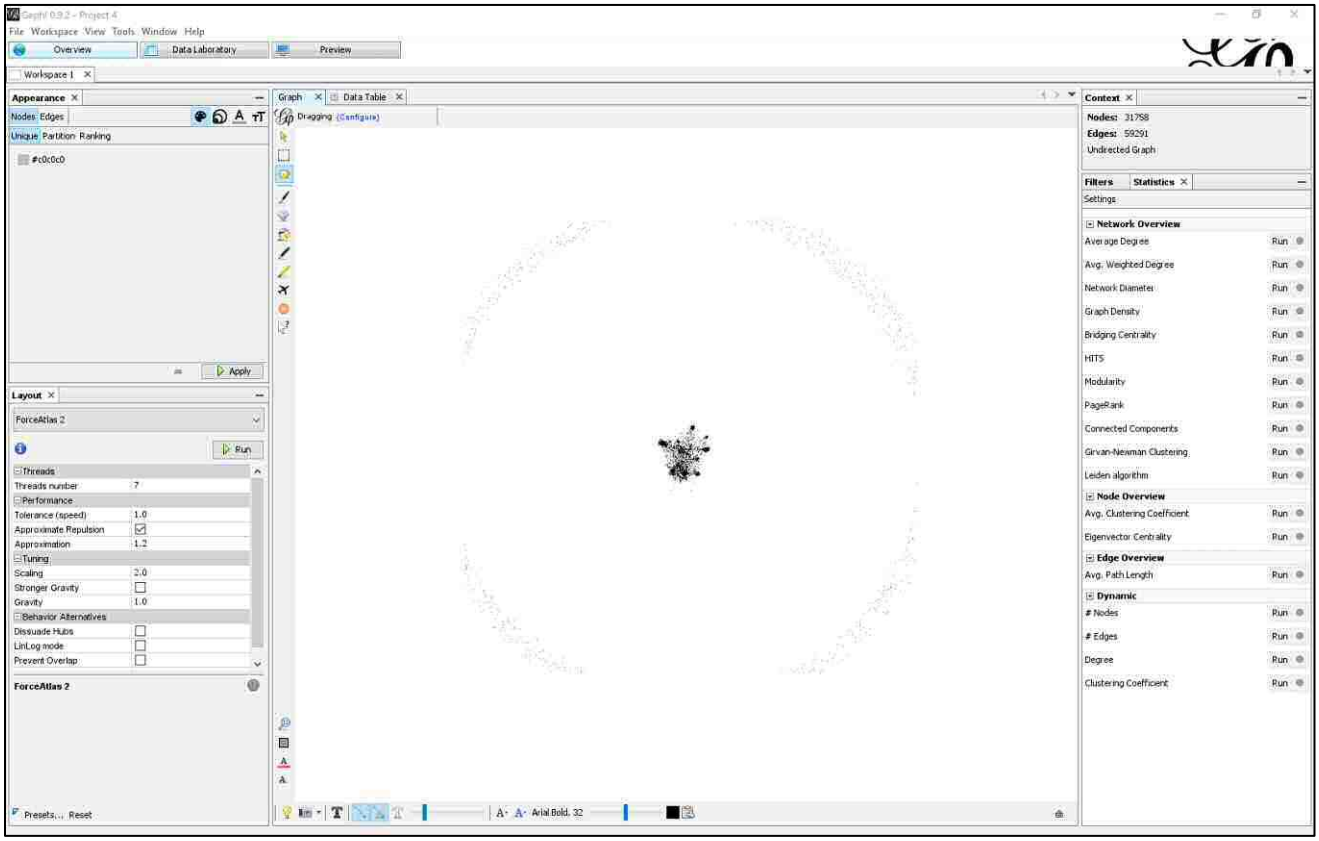

Figure 7. Graph after Force Atlas 2 layout applied.

The circle of nodes flung out to the edges of the graph represent the matches in the data set who are only matches with the autosomal DNA test used and have no other matches to anyone else – in other words, they are not "In Common With" matches. In the network representation, they have no edges with any other nodes and so are "repulsed" to the far outer edges of the graph. While individually they may certainly be matches worth pursuing especially if some share larger segments of DNA with the autosomal test used, this example will ignore them to focus on the clustered network at the center, which represents the matches which do have one or more matches among the others in the data.

Discussion

Characteristics of Network Clusters

At the center of this network graph, the nodes which do share edges with each other form into clusters, analogous to the squares in a Leeds Method analysis (this will be demonstrated shortly).

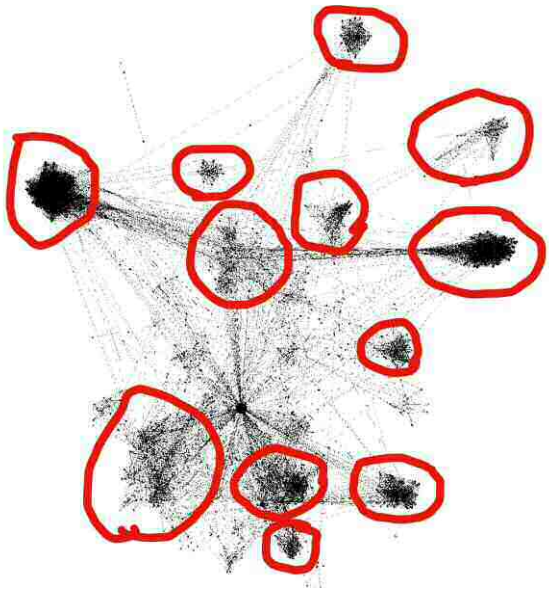

Figure 8. Nodes form clusters of various sizes and density.

Just as with a Leeds Method analysis, these clusters can be matched with groups who inherited common DNA segments from common ancestors. Clusters will vary in size and density according to the number of ICW matches, and because the Force Atlas 2 visualization algorithm causes nodes with higher interconnectivity to be attracted more strongly to each other, denser clusters represent subgroups of the network who have a higher incidence of ICW matching between each other than with the rest of the network. By definition, a subgroup of the network with a higher incidence of ICW matching will correspond to the descendants of common shared ancestors who share DNA from that same origin; however these common ancestors may be from a wide range of generations back in time, and some members of a cluster may be related more closely by later common ancestors who passed the same DNA along. So the shape and distribution of clusters will be very dependent on the closeness and degree of interrelationships between ICW matches for any given autosomal DNA tester.

Just as for Leeds Methods, clusters will also be less well-defined due to several influences, including:

1. Endogamy and/or pedigree collapse, which will cause clustering of matches who do share DNA from common ancestors but whose relationships will be from multiple common ancestral paths and not easily assignable to one common origin;
2. Pile-up regions of individual chromosomes, which will cause unrelated matches to have perceived ICW relationships that will show as connections between clusters on the graph, and may in significant cases blur the boundaries of otherwise unrelated clusters;
3. Very close relatives to the test being analyzed, who share DNA from several common lines of descent with the given autosomal tester and whose nodes will therefore show connections across many clusters.

All three of these influences can be somewhat mitigated by deeper analysis of the matches, shared DNA segments, and relationships, and changes to the network data to eliminate their influences. In particular, the third influence (close relatives) may be the easiest to identify and mitigate.

Changing the node sizes (in the Appearance Window) to reflect the amount of shared cMs, we can easily identify close relatives by node size. Figure 9 for example shows a portion of the network visualization where a close relative (a full sibling, in this case) has connections across various clusters.

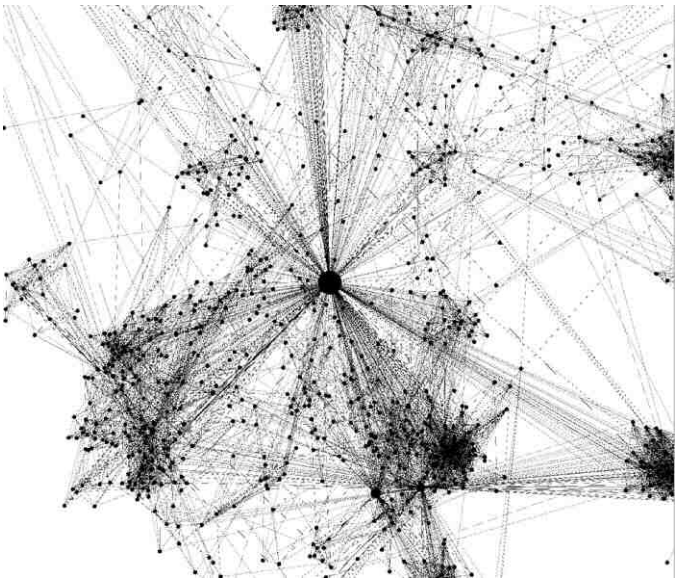

Figure 9. A subset of Figure 8 showing a full sibling crossing many clusters.

In some cases close relatives may be useful to retain; first cousins for example will generally show connections only to father's-side or mother's-side clusters which would aid the identification of origins. But for analysis of more distant clusters, their influence obscures the clear identification of clusters and close relatives can also simply be removed from the data set. For example, removing this full sibling from the data (and reapplying the Force Atlas 2 layout) would give us Figure 10, where cluster definition has significantly improved.

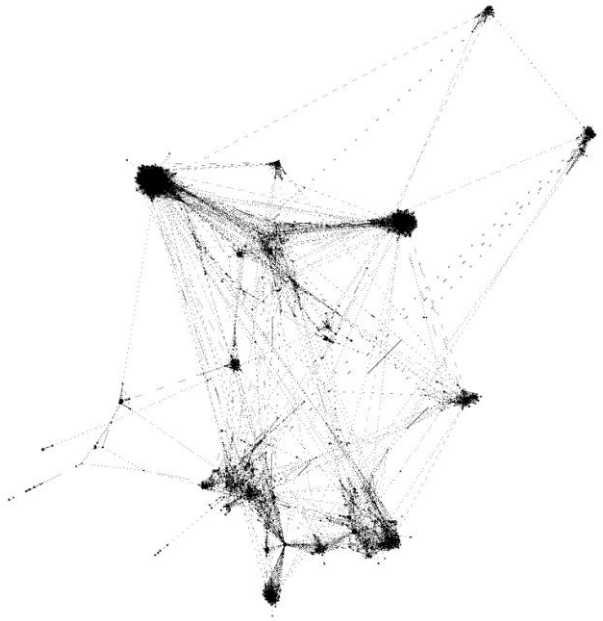

Figure 10. The network without the full sibling node.

Applying Color

The measure of the strength of division of a network into clusters is called *modularity*. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Gephi offers a calculation of modularity among its statistics functions in the right-hand window. The advantage of running these calculations is that they also create sub-components of the network which can then be colored in the Appearance window.

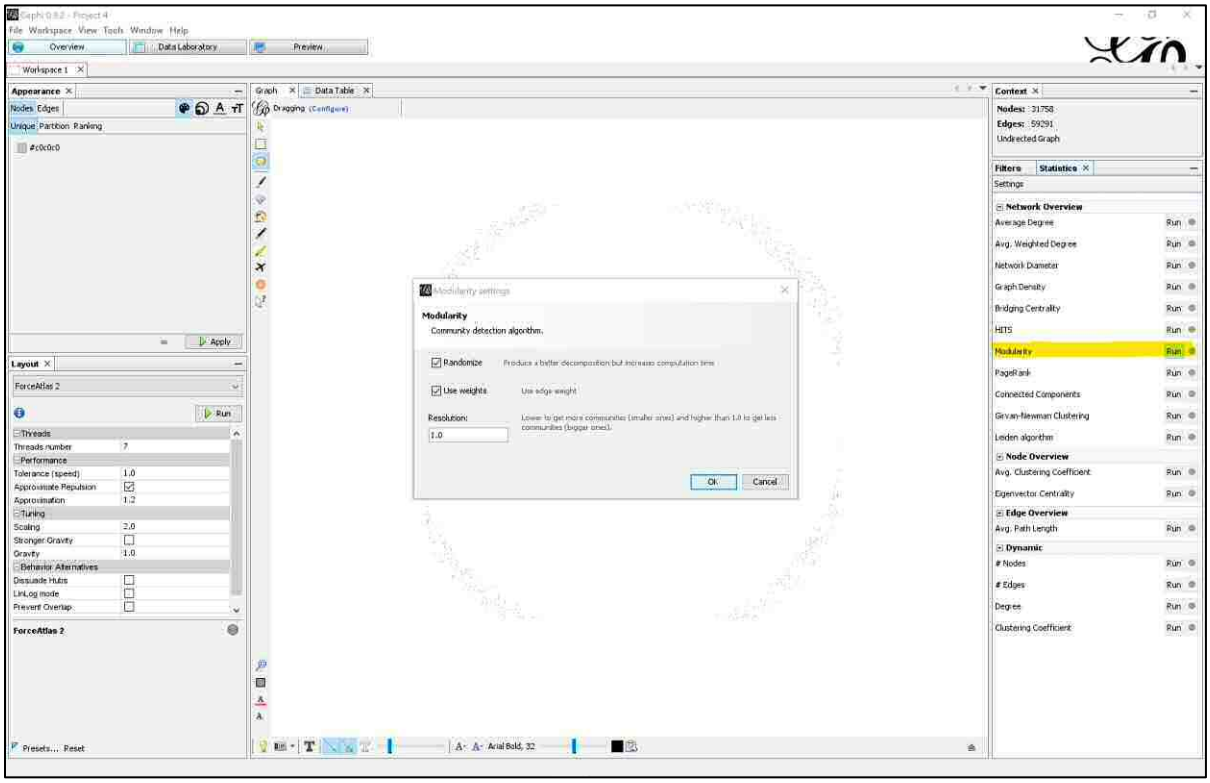

Figure 11. Running the Modularity Statistic

For the analysis of ICW matches, the "Randomize" option can be left checked although the "Use Weights" option is irrelevant since the edges are not weighted (although an extension of this approach might find it useful to weight them by shared cM or other additional information). The function also requires a "Resolution" number which drives the number of sub-networks that the modularity will determine. In order to color the visible clusters on the network, different resolution numbers may have to be used until the color groups match up as well as possible to the visible clusters. For this purpose a resolution of 3.0 seemed to work well for this data set, but it will vary somewhat according to the specific set of ICW matches.

Once the modularity function has completed it will produce a report, which for our purposes can be ignored. More usefully it will also have added a new column in the data called Modularity Class which labels the nodes by the divisions found by the modularity algorithm. In the Appearance window on the left, nodes can be colored by this Modularity Class. Before-and-after graphs of the central network of this coloring are shown in Figures 12 and 13.

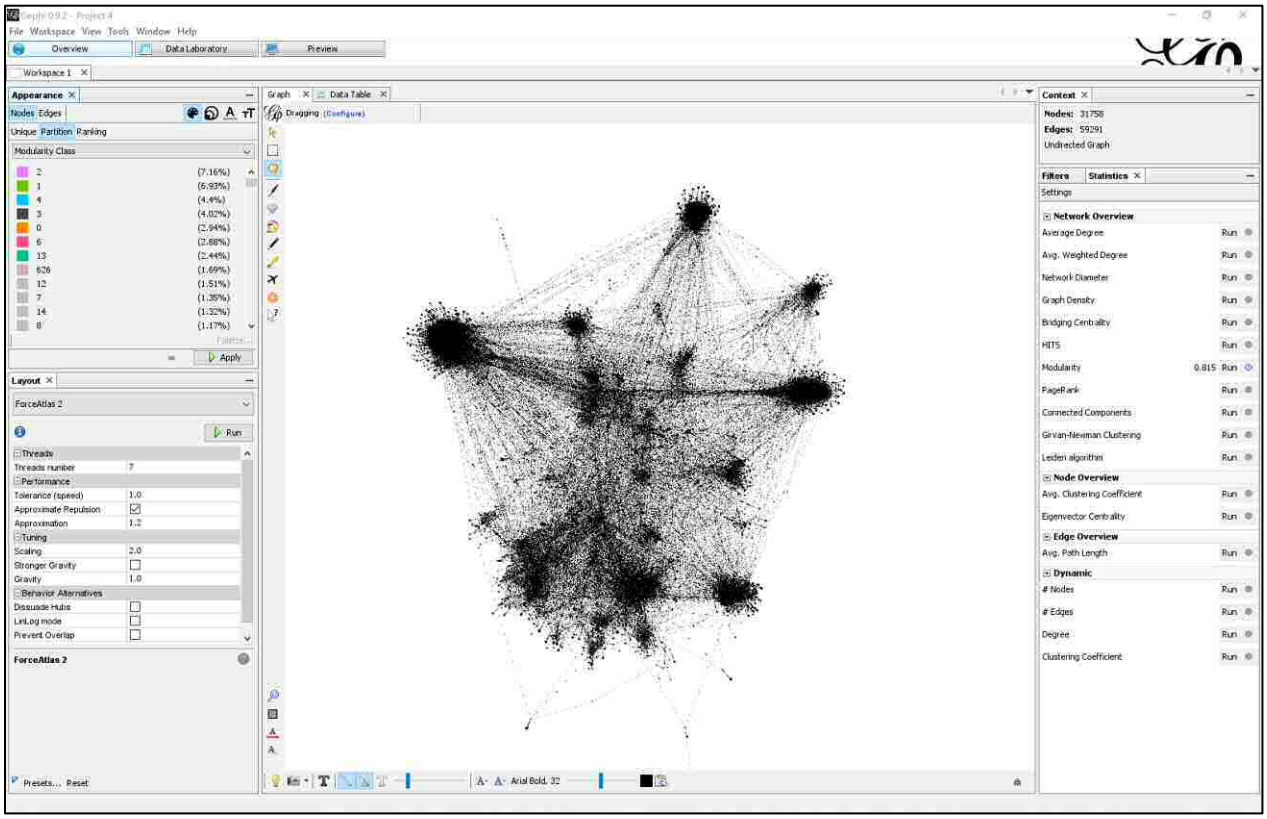

Figure 12. Before applying color by Modularity Class (see Appearance window in upper left).

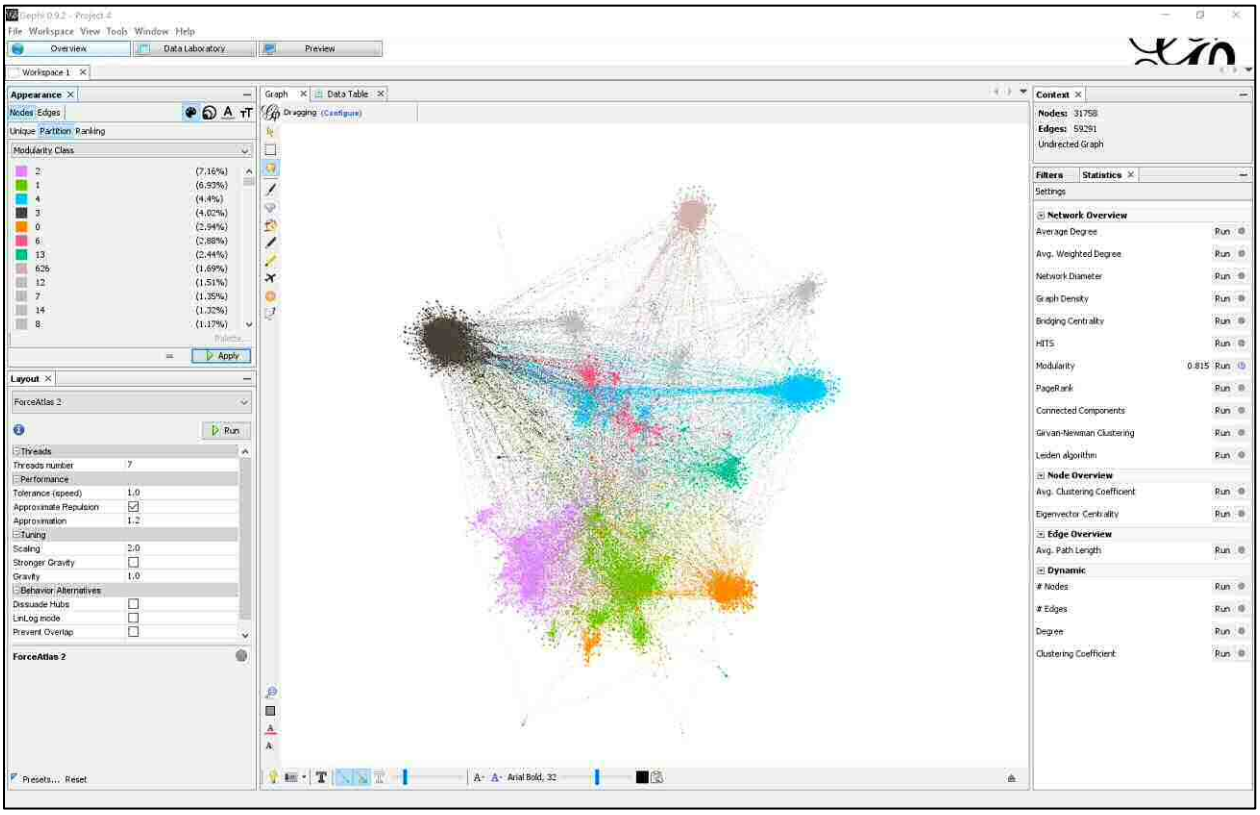

Figure 13. After applying color by Modularity Class

The Gephi tool allows color palettes to be automatically generated as well; although this feature is not covered in detail here, the application of the modularity statistic runs along with custom color palettes can be combined to shade the network in whatever number and type of colors will serve the analysis most usefully.

Labelling and Filtering the Graph

Another powerful feature of Gephi that is useful for this analysis is the ability to label nodes and display those labels on the graph. The Data Table has a column called "Label" which is initially blank but Gephi can copy data from one column to another in the Data Table.

Copying the match names into the Label column and turning labels on for the entire network is not particularly useful since it results in an unreadable picture (see Figure 14).

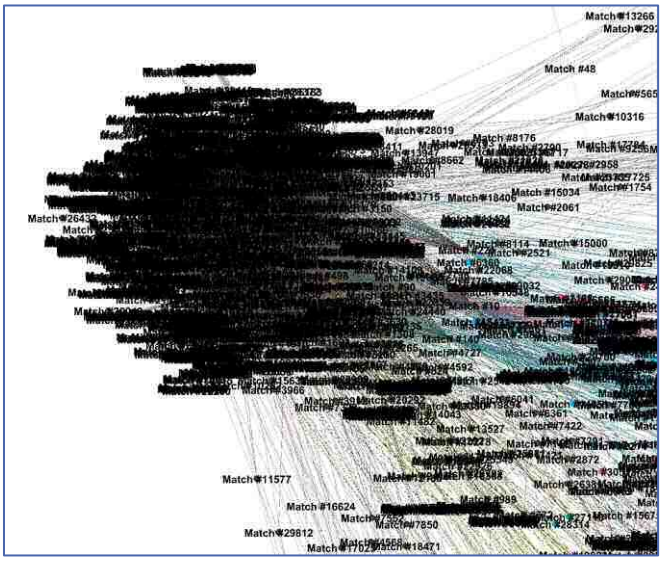

Figure 14. Part of the entire network with labels "on".

However, we can apply filtering to weed out unwanted detail and focus on subsets of the full network.

The Filter options in Gephi can be found on the right-hand side in the companion tab to the Statistics options. From here a variety of filters can be "dragged" down into the Queries window and applied to the network. For example, Figure 15 shows the full network with labels with a Range filter on Shared cMs dragged into the Queries window. This shows that the range of Shared cMs among the nodes ranges from 7.0 at the lower end up to 2741.52417 for the closest match. Changing the lower end of this range to 20.0 and applying this filter will eliminate any nodes (from the graph, not the data table) with less than 20.0 shared cMs . As Figure 16 shows, the display becomes much more readable.

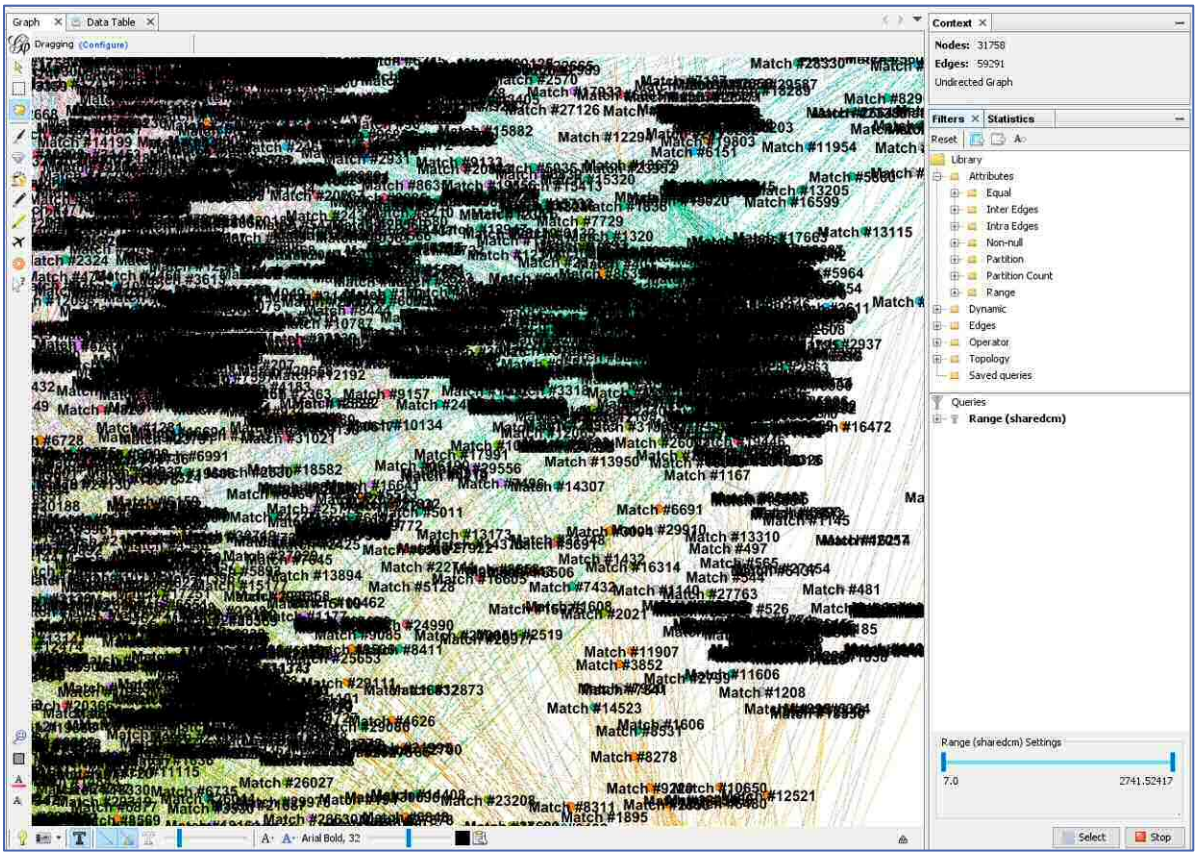

Figure 15. A Filter on Range of Shared cMs originally selected, showing the existing range among the nodes.

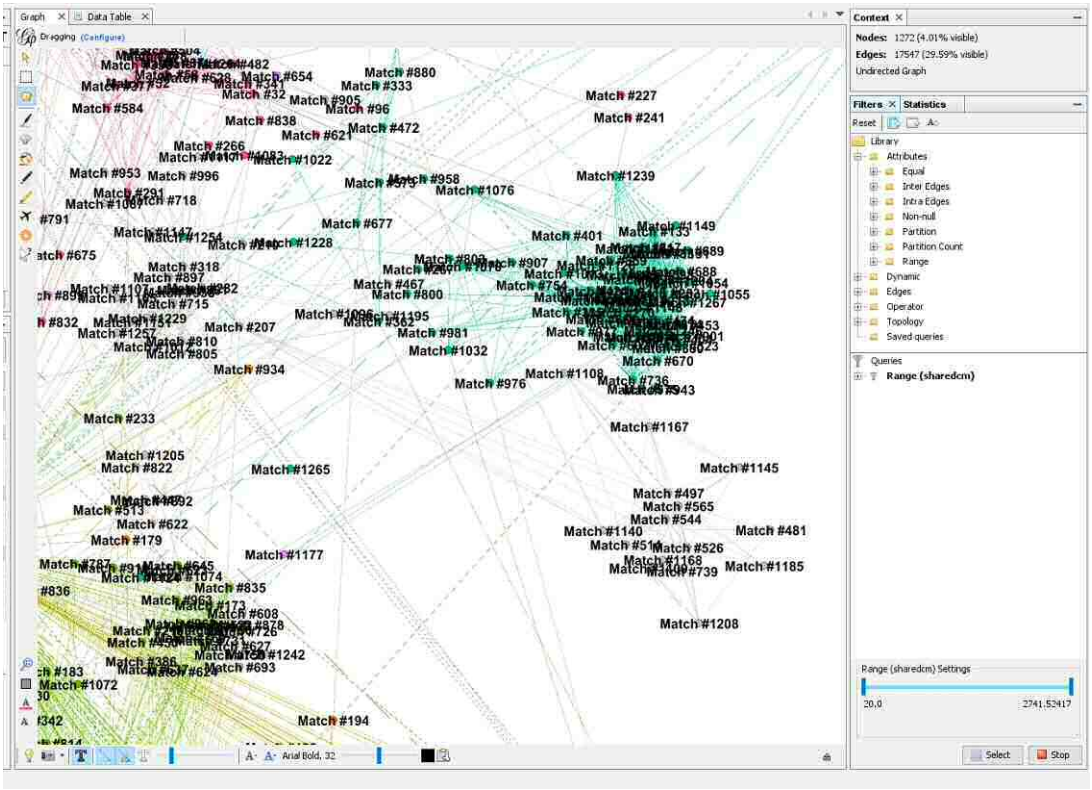

Figure 16. The graph after a filtering has been applied to limit the lower range to 20.0 shared cMs.

Exploring the graph with the nodes labeled by match name may be somewhat useful if clusters can be identified by common surnames or recognizable matches, but mapping other genetic and genealogical information on the network can gain us more insights.

In the Appearance window, we can also vary node size on the display by a data attribute like shared cMs. If we combine this with changing the node labels to the Known Relationships discussed earlier, we get the picture in Figure 17. Note that fewer nodes have labels since we only had Known Relationships for perhaps around 200 of the matches.

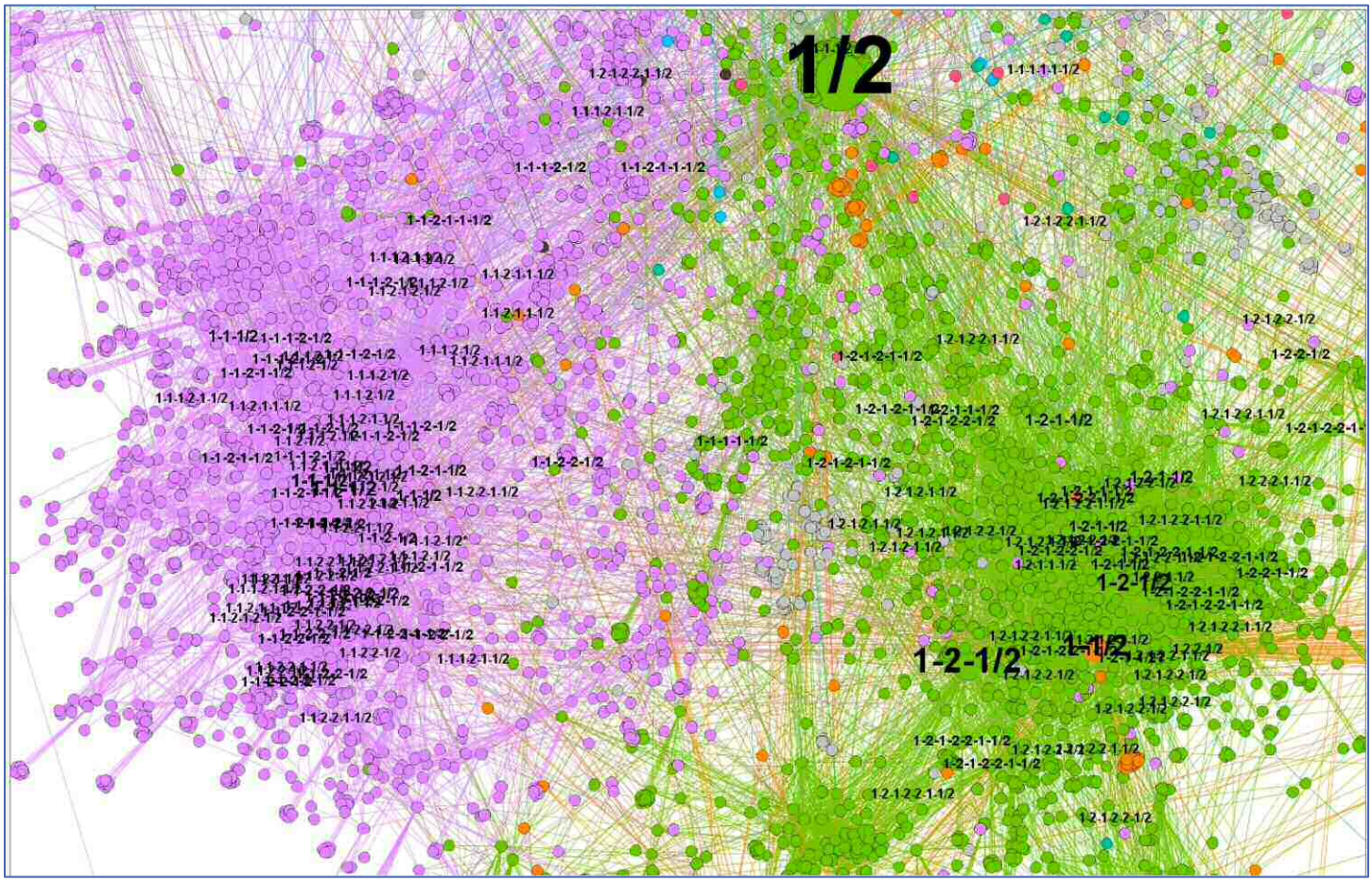

Figure 17. A portion of the network with node and label sizes by shared cMs and Known Relationships used as labels.

Figure 17 is still confusing both from the number of nodes and the influence of the full sibling node at the top of the figure. Removing this sibling node and filtering to eliminate matches in the 7 to 20 cMs range of shared DNA gives us Figure 18:

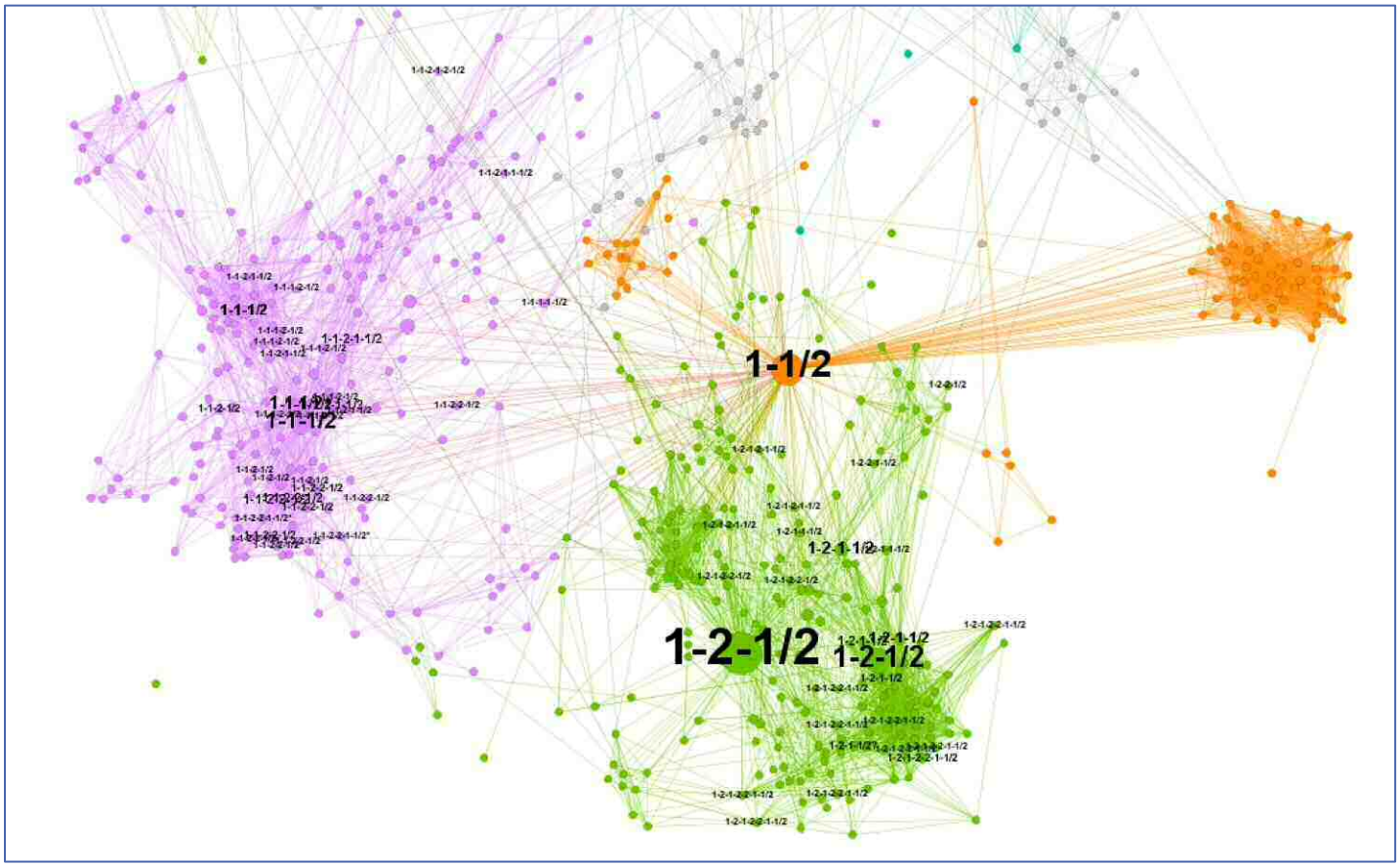

Figure 18. Clusters of nodes above 20 cMs without the full sibling node.

While we could continue to remove the full cousin and full second cousin nodes shown in Figure 18, the origin of these clusters is relatively clear. The purple cluster on the left appears to originate from at least the 1-1-x side of the author's ancestry, meaning our paternal grandfather. There are known relationships on both the 1-1-1-x and 1-1-2-x ancestral lines showing in the purple cluster; these might show better cluster definition if the closer relatives were removed, or perhaps the shared DNA is from both sides. The green cluster however appears to connect in our paternal grandmother's ancestry (i.e. 1-2-x), with a heavy bias to the 1-2-1-2-2 ancestral branch. Again eliminating closer nodes might help break up these clusters for analysis.

The other orange cluster in Figure 18 shows no Known Relationships, but a very tight connection among these ICW matches. Given their consistent connection to the first cousin marked as "1-1/2", they clearly connect on our father's-side ancestry but more cannot yet be determined. After finding the match names associated with these nodes, it is likely that any progress on identifying the common ancestor with any of these matches will narrow down the connections for all of them.

The first cousin's (1-1/2 node) multiple connections to the orange cluster is a pattern that can show with many clusters in the network. Figure 19 shows another cluster with two matches with known relationships to the 2-1-2-1-x ancestral path (and with shared cMs of 51 and 56 cMs), connecting to a

cluster formed of nodes with cMs between 20 and 27 cMs.

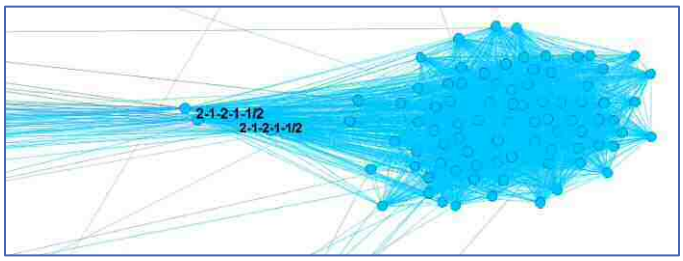

Figure 19. Another cluster in the network

While again endogamy, pedigree collapse and pile-up regions can cloud or even resemble this pattern, it often also is due to closer relations with larger cMs who have inherited DNA from the same older ancestry as the cluster – i.e. that this cluster in Figure 19 has a common ancestor in the generations before our 2-1-2-1 ancestor's parents (in other words, either our mother's father's mother's father's father or mother). Narrowing down these matches at least this far in our common ancestry should significantly reduce the necessary research to identify our specific common ancestors.

In some cases of course, the common ancestry with these matches which connect to a larger cluster will not be known. In those cases these nodes can help focus our research priorities, since identifying the ancestral line of these "connecting nodes" means that the matches in the cluster are also likely connected by ancestors along the same ancestral line.

Making Charts to Share with Family

Separate from analysis (or sometimes to help explain it to others), Gephi can also produce good quality network charts like a simple example in Figure 20 which (if well-explained) can help explain the results of autosomal DNA analysis to other family members as well.

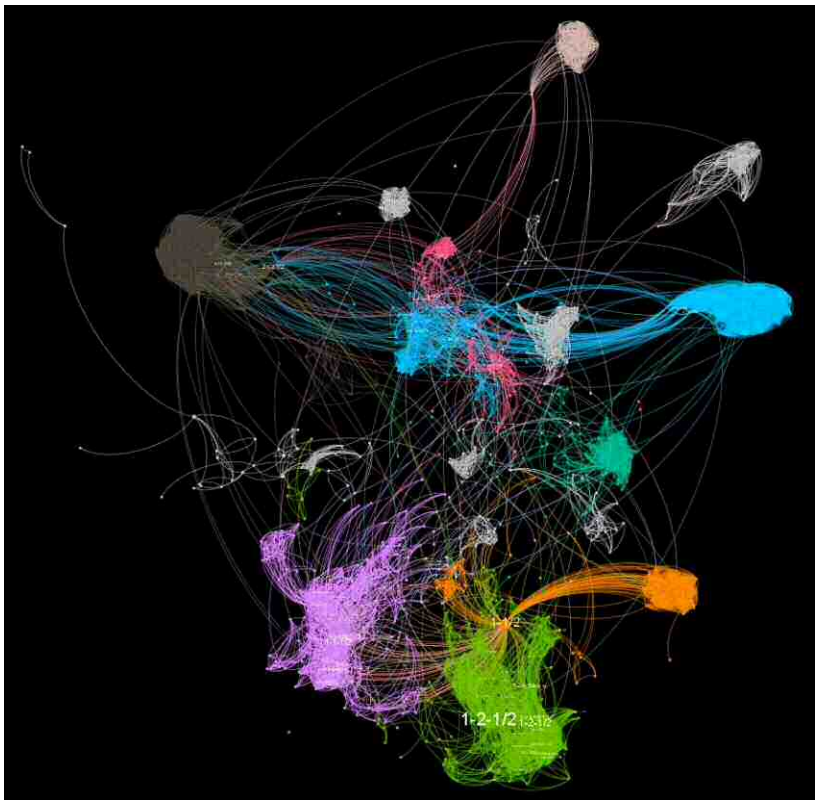

Figure 20. Small changes in color and display can result in shareable views

Comparison to Other Methods

The Collins-Leeds Method (CLM) analysis mentioned earlier was used to generate cluster squares organizing 2,870 of the matches. This was done solely for purposes of this review to provide a comparison of Collins-Leeds Method clustering with Force Atlas 2 clustering.

To perform this comparison, the CLM square numbers for the 2,870 nodes were copied into the Nodes Table data set. Changing the Appearance window to color the nodes based on this column then provides a visual comparison between the clustering approaches in Figure 21. Grey nodes in this figure represent matches who were not sorted into CLM clusters and therefore have no number in the clm_cluster column.

The comparison performed was only by visual inspection since this was performed only to show general alignment and exact correlation was not required. While this is a subjective comparison and not a statistical one, it is clear from Figure 21 that the CLM squares align well with clusters produced by the Gephi analysis.

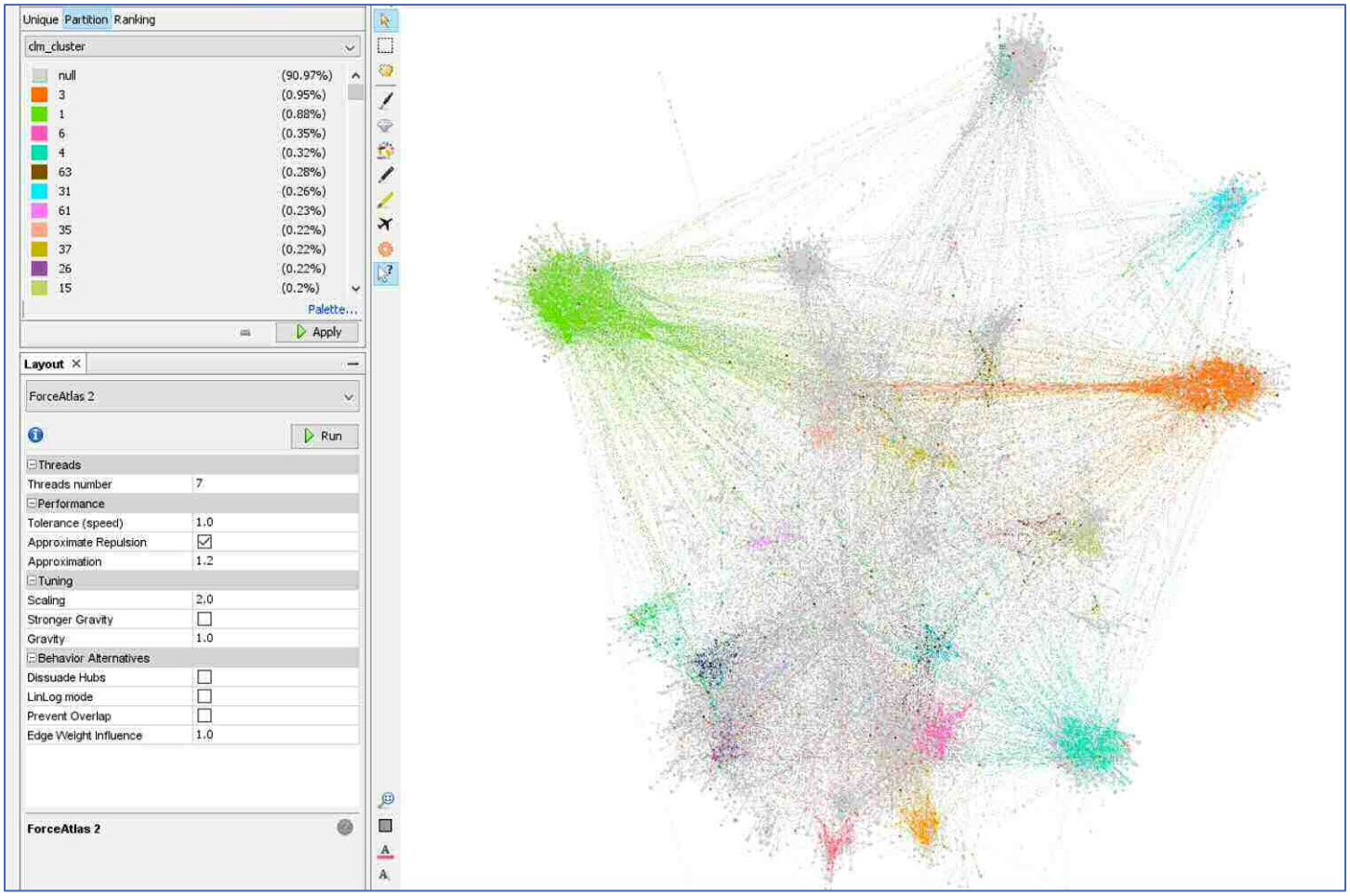

Figure 21. Coloring the CLM Clusters demonstrates a 1-to-1 equivalence with network clusters.

Splitting the Network into Paternal and Maternal Matches

In theoretical models of ancestry, inherited DNA is cleanly split between the portion inherited from a father versus from a mother. While the occurrence of matches will be influenced more by number of children along ancestral lines and the somewhat random chance of lines surviving to present day and being tested, one would still assume that a network of ICW matches would break cleanly into two unconnected super-sets of clusters representing our separate paternal and maternal DNA ancestry. In practice, the split is rarely that clean given the confounding influences mentioned previously. However, by changing the modularity filter's resolution until only two colors remain, we can at least see how easily the network might split into two sub-networks.

To illustrate this, Figure 22 shows the network after the modularity filter has been run with a large enough resolution that only two major colors appear (in this example a resolution of 20.7 was used).

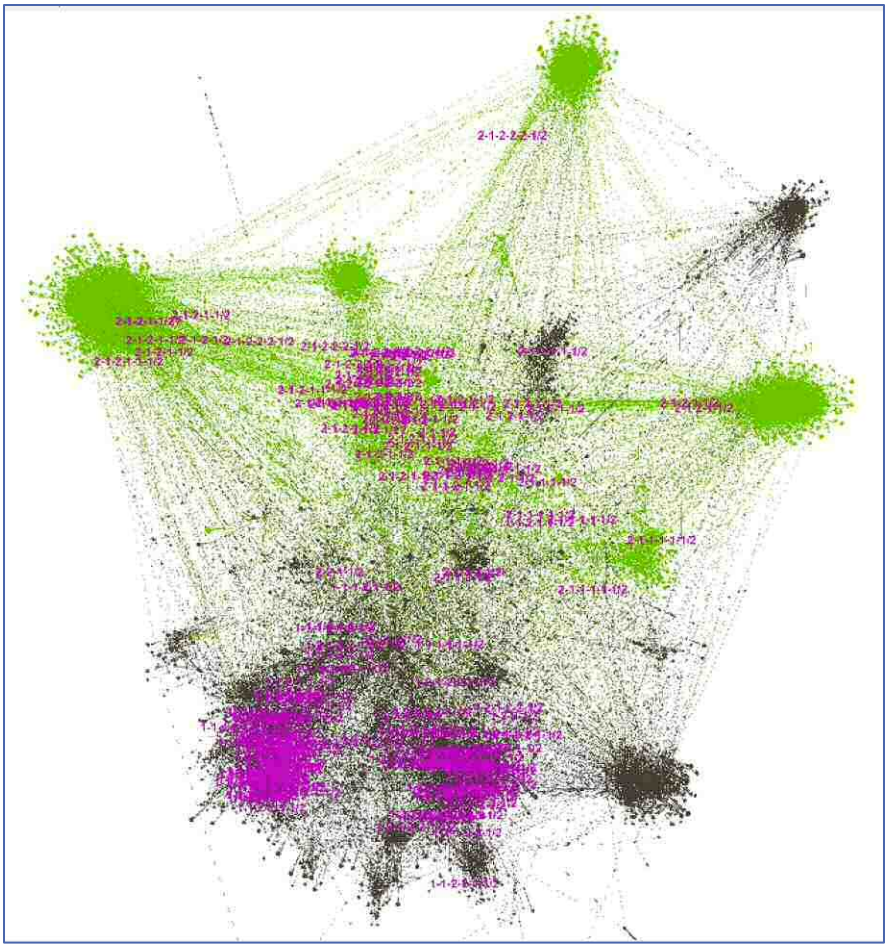

Figure 22. Splitting the Network into two super-sets (in this case, green and black)

For this purpose network visualization is particularly powerful since the identification of these super-sets does not require any knowledge of whether certain matches are paternal or maternal although without any knowledge we could still not assign the super-sets to one side of our ancestry or the other. In Figure 22 however we have increased the size on the Known Relationship labels to look at this more closely. In Figures 23 and 24 the close-ups of Figure 22 show that for our known matches, the green nodes are more clearly from our maternal side and the black nodes from our paternal side.

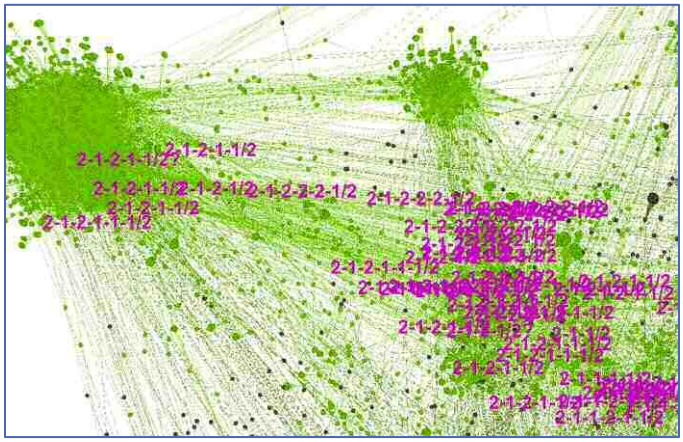

Figure 23. The green nodes are from our maternal side (2-x)

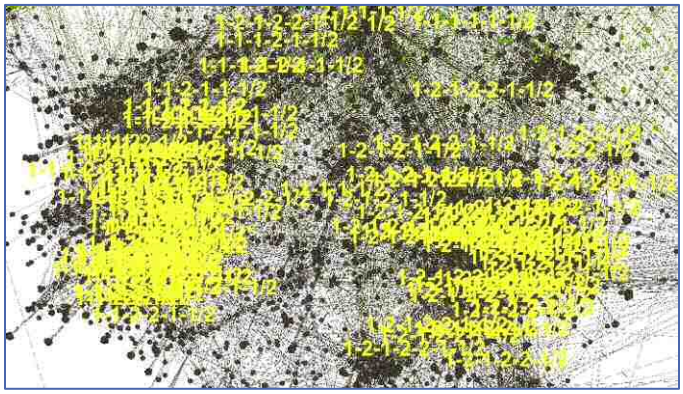

Figure 24. The black nodes are from our paternal side (1-x). (Label colors changed for clarity)

Note that this approach may inadvertently pull some matches from one side over to the other because of connecting edges that confuse the analysis, so especially for nodes that do not aggregate cleanly into large clusters, this split should be taken as suggestive, not conclusive for all nodes.

Looking more closely at the inhibitors to whether a network of ICW matches will split into paternal and maternal relatives, the largest influences may be from the levels of shared endogamy and/or pedigree collapse across the tester's paternal and maternal ancestries since higher levels of either will cause more significant cross-connections that would inhibit the identification of separate sub-networks. DNA pile-up regions may also cause some cross-connectivity between clusters. If the tester's full siblings were also present in the network their nodes would also show significant cross-connections to both paternal and maternal sub-networks; however these nodes can be relatively easily identified and eliminated. Half-siblings and any degree of cousins in the network would not inhibit the identification of paternal and maternal sub-networks since they would (at least based on their primary relationship to the tester) only have connections to one sub-network and not the other.

However even if based only on the potential for endogamy and/or pedigree collapse, not every ICW network will be this easily divisible into paternal and maternal sub-networks. In some cases this may only provide general clues as to whether an unknown cluster is more likely to be from a tester's paternal or maternal sides, but in other cases the exercise may be inconclusive.

If the exercise does identify paternal and maternal sub-networks, one could envision repeating the exercise on each subnetwork to further sub-divide each into grandparent sub-networks and even further. It is however unlikely that any tester's ICW network will subdivide cleanly over several generations, and for that analysis half-siblings or cousins who share ancestry back to those generations would definitely need to be removed from the network.

It is also possible that a similar approach combined with some knowledge of the known origins of specific clusters could be used to identify areas of the network which were affected by endogamy and/or pedigree collapse and so could localize instances of either in the tester's ancestry. This has not yet been studied but remains a potential extension of network visualization analysis.

Extending the Analysis

This review only covers some approaches for using network visualization to focus the likely ancestral lines shared by match clusters to assist in identifying common ancestors. These approaches could easily be extended; for example to include more detailed information from Leeds Method analysis, or details of shared segments by chromosome; or different filtering options could be applied to restrict the network to ranges of shared cMs that correspond to specific limits of generational differences. Once an analyst has some familiarity with the platform used to visualize and filter the network, it can provide a myriad of analysis techniques along with real-time flexibility to switch between them.

Supplementary Information

The author has a video demonstrating these techniques along with more introductory explanations and more detail on working with Gephi options. This video can be found on YouTube at https://youtu.be/Z2T_7aSL4ng.

Conflicts of Interest

The author declares no conflicts of interest and no commercial interest in the topics covered in this review.

References

Gephi tool website, https://gephi.org/. Accessed October 7, 2021.

DNAGedcom tool website, https://www.dnagedcom.com/. Accessed October 7, 2021.

DNAPainter tool website, https://dnapainter.com/. Accessed October 7, 2021.

Genetic Affairs tool website, https://geneticaffairs.com/. Accessed October 7, 2021.

Plos One, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". https://doi.org/10.1371/journal.pone.0098679. Accessed October 7, 2021.

Nicholson, Brit. "Auto-Clusters in Gephi Using Data from GEDmatch". https://dna-sci.com/2020/08/19/auto-clusters-in-gephi-using-data-from-gedmatch/. Accessed October 7, 2021.

YouTube video, "Autosomal ICW Match Analysis using the Gephi tool", https://youtu.be/Z2T_7aSL4ng. Accessed October 7, 2021.