Faculty of Natural Sciences and Technology
Department of Physics

NTNU
Norwegian University of
Science and Technology

MASTER'S THESIS FOR

STUD. TECHN. JON WOSTRYCK EIESLAND

Thesis started:	20th of January 2009
Thesis submitted:	20th of July 2009

DISCIPLINE: THEORETICAL PHYSICS

English title:	*Communities in a large social network - visualization and analysis*
Norsk tittel:	*Samfunn i et stort sosialt nettverk - visualisering og analyse*

This work has been carried out at Helsinki University of Technology, Finland, under the supervision of Professor Kimmo Kaski and Dr. Tech. Jari Saramäki.

Responsible supervisor at NTNU:

Alex Hansen

Professor at the Department of Physics

Abstract

Communities have been a hot topic in complex network research the last years. Several algorithms for detecting communities have been developed, and in this thesis we use the sequential clique percolation algorithm to detect communities in a large social network. Our network consists of 5.3 million mobile phone users, with mutual communication data aggregated over 18 weeks.

In this thesis we do a visual study of the communities, and we clearly see the nested community structure when we do clique percolation for different clique sizes. When we threshold the edge weights we see that the strongest edges are in the densest sub-communities and that the weakest edges keep the communities connected.

We also present numerical analysis of some selected structure and topology properties of the communities. Lastly we confirm, by numerical analysis of the available demographic data on the mobile phone users, that the communities are more conform with respect to zip code, age and sex compared to a reference network where the demographic attributes have been shuffled.

Sammendrag

Samfunn har vært et hett emne innen forskning på komplekse nettverk de siste årene. Det har blitt utviklet flere algoritmer for å finne samfunn, og i denne oppgaven bruker vi sekvensiell klikkperkolasjon til å finne samfunn i et stort sosialt nettverk. Nettverket vårt består av 5.3 millioner mobiltelefonbrukere, med gjensidig kommunikasjonsdata aggregert over 18 uker.

I denne oppgaven gjør vi en visuell studie av samfunnene, og vi ser tydelig den vevde sammfunnsstrukturen når vi utfører klikkperkolasjon for ulike klikkstørrelser. Når vi setter terskler for lenkevektene ser vi at de sterkeste lenkene er i de tetteste undersamfunnene og at de svakeste lenkene holder samfunnene i kontakt med hverandre.

Vi presenterer også en numerisk analyse av noen utvalgte struktur- og topologiegenskaper hos samfunnene. Til slutt bekrefter vi, via numerisk analyse av den tilgjengelige demografiske informasjonen om mobiltelefonbrukerne, at samfunnene er mer konforme med tanke på postkode, alder og kjønn sammenlignet med et referansenettverk hvor de demografiske attributtene har blitt stokket om.

Preface

This work was done at the Department of Biomedical Engineering and Computational Science (BECS) at the Helsinki University of Technology, Finland (TKK). I would like to thank Professor Alex Hansen, my supervisor at NTNU, and Professor Kimmo Kaski, head of the Center of Excellence in Computational Engineering at BECS, for giving me the opportunity to work there. Their support and encouragement has been very valuable to me. I would also like to thank Dr. Tech. Jari Saramäki for overseeing this work and giving highly inspired feedback throughout the process of writing. The rest of the staff at BECS has also been very friendly and helpful and I have greatly enjoyed my stay with them.

Finally, I thank my mom and dad for their wholehearted support and my brothers for all the great times we have had.

Trondheim, 9th of July 2009
Jon Wostryck Eiesland

Figure (1): This comic strip nicely sums up the joys (or pitfalls?) of doing computational work. Courtesy of http://xkcd.com/, under a Creative Commons Attribution-NonCommercial 2.5 License as of June 2009.

Contents

Chapter 1

Introduction

The interest in complex social networks research has greatly increased in the last decade as the increase of computational power has made it possible to analyse ever larger networks. The development of new digital channels for communication and information flow has further made it feasible to collect enormous sets of precise data regarding human relationships. The largest network studied so far can easily be called "planetary"[1], as it looks at the communication patterns of 180 million people around the world using an instant messaging service [1]. This is quite a leap from the study by Zachary, of the splitting of a karate club with 34 members [2]. The karate club has been extensively studied by sociologists since it was published in 1977.

Lately, the study of communities has been one of the focal points of complex network research. Communities are interesting as they often correspond to the functional groups of a system. Much effort has been put into the development of models and algorithms for searching out the denser connected groups of nodes in a network, i.e., communities. The concept of a community can take different shapes in different networks, however, meaning that the community structure found is very much dependent on the algorithm used. Both global optimization of communities and local topology only, are valid starting points for community detecting methods. In social networks one wants to be able to detect communities consisting of family, friends, coworkers and so on. Separating these communities from each other can be practically impossible without extremely detailed information about the persons in the network, though. In this thesis we will use a specific clique percolation algorithm to detect the communities in our mobile phone user network. We have chosen this model because of its logical definition of communities, its simple interpretation, its speed and its high suitability to visualization.

The ever growing size of analysed networks hampers the possibility to make a meaningful visual representation of them. A good representation of the subject under study is something one should never underestimate, though, and the increasing computational power can to some extent be put to good use also in this area. Presenting an acceptable

[1] "Global", however, would not be a correct term to use. The use of instant messaging is not common in several highly populated parts of the world like many African countries and rural China, quietly excluding them from the data.

layout of a large network in two dimensions is no longer a problem, but for the time being, computers can neither improve the limited resolution of the human eye (save for some on-screen zooming of course) or print out three-dimensional plots. One of the goals of this thesis is therefore to put one of the many community detection models to action on a large social network and present a good visualization of the resulting community structure. We also want to learn more about the behaviour of communities by quantifying a selection of their network properties. Since it is impossible to get such quantifiable information from a visualization, we have also done numerical analysis on both structural and demographic properties of the communities.

When one has taken the step up to do analysis of networks consisting of millions of persons, the individuals can not have the same focus anymore. We are forced to use statistics to present the average behaviour of people. This may seem alarming, but we will see that several properties of our network have so broad distributions and large variances that the individuality of human behaviour is not simply discarded.

1.1 The layout of this thesis

In chapter 2 we will first give an introduction to complex networks and communities. Selected properties and interesting topics on the subjects will be discussed. We also describe the algorithm we have used in this work to detect communities, sequential clique percolation. The large social network we have been working with in this thesis comes from mobile phone user data from a mobile phone subscription company. In chapter 3 we describe this data and how we treat it in a network model.

In chapter 4 we present one of the first visual studies of communities in a large social network. We have plotted the communities detected in four of the largest groups of nodes in our network and also incorporated the demographic data available to us. To fully understand the communities one also needs to do numerical analysis, and in chapter 5 we have analyzed certain properties of the detected communities. First we look at the pure structure and topology of the network and then we take a look at the demographic composition of the communities.

Chapter 2

About complex networks and communities

2.1 Complex networks

A *network* (or *graph*) is mathematically defined as a pair $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of *nodes* (or *vertices*) connected by a set of *edges* (or *links*) $\mathbf{E}$. Any kind of items, organisms or abstractions which have relations between them can be represented by a network model.

In most networks an edge $e \in \mathbf{E}$ may connect only two nodes $i, j \in \mathbf{V}$, and may be written as $e_k = (i, j)$. The nodes connected by e are *neighbours*. In a *directed network* the edges e_{ij} and e_{ji} represent two different objects. One is the relation from i to j and the other in the opposite direction. E.g., the original phone call data we are using in this work, has a clear difference between caller and receiver, or a food web where species are nodes and an edge (i, j) means that species i eats species j. In an *undirected network*, the direction of relations is not taken into account, or may not even give any meaning. E.g., scientific collaboration networks are undirected: each node being one person, and an edge $(i, j) = (j, i)$ means that i and j have collaborated on some work.

We use $N = |\mathbf{V}|$ for the number of nodes and $L = |\mathbf{E}|$ for the number of edges. Assuming no self-loops (edges on the form (i, i)), the maximum number of edges in a network is $N(N-1)/2$ for an undirected network (a situation where all the nodes are connected to all the other nodes), and the double for a directed network (one edge in each direction). The *edge density* is just the actual number of edges divided by the possible number of edges.

What exactly does it mean that a network is "complex"? A small note on this subject is required. First of all, complex networks distinguish one field of research from many others dealing with or making use of graphs and networks. Secondly, the term points out that the dynamics of all the components in the network and the emergence of certain properties might not at all be obvious. A network of millions of nodes and edges is no more complicated than an airplane with millions of parts itself, but while the airplane and its components work in a well known way, the inner workings and dynamics of the network

might be very hard to predict (see, e.g., the small-world effect in section 2.1.2). Social networks, in particular, are very difficult to understand by their components alone. People think and act in peculiar ways, and we have no possibility to handle all this information if we even had it.

That was big words and it is tempting to fall back on the more casual explanation: complex networks are big, millions of nodes and edges, but the number of edges is generally small compared to the possible number of edges. Furthermore, the structure of interactions depicted by the edges is non-trivial, it is far from regular, but not entirely random either.

2.1.1 Weighted and dynamic networks

Edges in a *weighted network* have a property called *weight*. This property usually gives a measure (again usually, an integer or a real number) of the strength of the relationship between two nodes. A plain edge in an *unweighted network* simply describes a generic relationship. In many cases, introducing weights to a network model brings more realism and definitely more opportunities for analysis. Take your average social network for instance, people are closer connected to their families than to their friends, and one has close and distant friends. Setting weights can be tricky unless one has a connection between nodes which one can count and aggregate over a time period. For example, we use the number of phone calls between people in our network, and the number of passengers transported can be used in an airport network. One may also want to do analysis based on different weights in the same network, e.g. calls versus SMS or passengers versus flights.

The next natural point of discussion is dynamical networks. Many networks obviously evolve with time, they are dynamical. An unweighted network may evolve with the addition of new nodes and links and the removal of old ones. In a weighted network one might have the additional growing and shrinking of weights. The normal procedure for analysis of a network is that one just take a snapshot of it at a certain point in time. With counted weights you need a reference point to start counting from, of course. We will later use calls between persons aggregated over half a year as weights in our network. As long as there are no sudden global effects happening in the network to take into account and your net is large enough, this method will give rise to some fine analysis. There is also the possibility that the network one is looking at has reached a stable configuration. This is in the sense that, sure, individual nodes have come and gone, but the overall structure and properties of the network are more or less constant. This has been shown with an US airport network over ten years in [3]. One could assume that it is also the case for any large social network, with subtle changes with the ease of communication by new technology, i.e., telephone, airplanes and the Internet. Researchers are currently experimenting with shorter time scales. For example, on-line information flow has been studied and a “distance” in the network has been formulated by the shortest time it takes for information to transfer between two nodes [4].

2.1.2 Additional definitions and properties of networks

Degree and strength

In an undirected network, a node's *degree* is the number of edges connected to it. The average degree of a network is, since each edge has two endpoints,

$$\bar{k} = \frac{2|\mathbf{E}|}{|\mathbf{V}|} = \frac{2L}{N}. \tag{2.1}$$

The *strength* of a node is the sum of the weights of the connected edges

$$s_i = \sum_j w_{ij}. \tag{2.2}$$

In a directed network, the nodes will have both an in and an out value for degree and strength.

One of the first points in any network analysis is usually to plot the degree distribution. If the fraction of the nodes in the network with degree k is p_k, one can plot p_k for a network by making a histogram of the degrees of the nodes. The problem is, however, that degrees and many other properties of networks often have very *fat-tailed* distributions. Fat-tailed meaning, in broad terms, that a large variance in the distribution is caused by a few extreme deviations, as opposed to a great number of small deviations in a distribution resembling the bell-shaped Gaussian curve. Real-world networks often follow power-laws in the tail of their distributions, meaning that they have a long right tail with nodes with degree far higher than the average. This was one of the focal points in early complex network research [5]. The power law being on the form $P(k) \sim k^{-\alpha}, \alpha > 1$, with the cumulative distribution $P(K > k) \sim x^{1-\alpha}$. If $1 < \alpha < 2$ both the mean and the variance of the distribution is infinite, if $2 < \alpha < 3$ the mean is finite and the variance infinite (many real-world networks being in this range), and if $\alpha > 3$ both are finite. The exponential distribution, $P(k) \sim e^{-k/\kappa}$, with the same exponent in the cumulative distribution, is also common to bee seen.

Getting a good measurement of the distribution's tail and making the mentioned plot can be very tricky. The direct histogram approach fails since one rarely has enough samples in the tail to get good statistics. There are two accepted ways to deal with this problem. One is to construct a histogram where the bins' sizes increase exponentially with degree. The number of samples in each bin is then divided by the bin width to normalize the measurement. Thus the bins appear even sized when plotted in logarithmic scale. This is how some of the plots in this work have been created. Another way is to plot the cumulative distribution function. Such a plot has the advantage that it preserves all the information from the data. When we make a binning histogram the differences between data points falling in the same bin are lost. However, the cumulative plot does not give an actual visualization of the data and the adjacent points on the plot are not statistically independent, making it difficult to correctly fit the data.

Power-law and exponential distributions can be spotted experimentally by plotting the cumulative distributions with logarithmic scales (power-law) or semi-logarithmic scales

(exponential) and looking for straight lines. However, other distributions (e.g. lognormal and stretched exponential) can disguise themselves to look like straight lines when plotted and noise can also be a factor. In general, the naked eye can not have the final word.

Clustering coefficient and assortativity

A clear difference between real-world networks and random graphs is the property of *clustering*. In many real-world networks we see that if there is a connection between node A and node B and between B and node C, then there is a heightened probability that A and C are also connected. In the language of social networks one would say: the friend of your friend, is likely to also be your friend. So we have triangles in the network. One way to mathematically define the clustering coefficient of node i is

$$c_i = \frac{t_i}{k_i(k_i-1)/2}, \tag{2.3}$$

where t_i is the number of edges within i's neighbours, and the denominator the possible number of such edges. One does not have a defined c for nodes with $k = 1$. In most observed networks, the clustering is higher than in a random reference network of the same size [6].

In most social networks it is shown that nodes with larger degree tend to be connected, they are said to be assortative. People with a lot of contacts tend to know other people with many contacts. The common test is whether the average nearest-neighbour degree increases or not with degree of the nodes. The opposite is true for many other networks (e.g. technical, information and biological). This is interesting, since degree is a property of the network topology itself.

Intensity and coherence [7]

The intensity $I(g)$ of a sub-graph g with nodes v_g and edges l_g, such that the number of links in g is $|l_g|$, is defined as

$$I(g) = \left(\prod_{(ij)\in l_g} w_{ij} \right)^{1/|l_g|}, \tag{2.4}$$

where w_{ij} is the weight of (i, j). It is basically the geometric mean of the edges' weights. Intensity measures the "weight" of a sub-graph, and it can be compared to the other sub-graphs in a network. Due to the nature of the geometric mean, the intensity can be low because one of the weights is very low or because all the weights are low. To distinguish between these two extremes, sub-graph coherence $Q(g)$ has been introduced as the ratio of the geometric to the arithmetic mean of the weights

$$Q(g) = \frac{I(g)}{1/|l_g| \sum_{(ij)\in l_g} w_{ij}}. \tag{2.5}$$

$Q \in [0, 1]$ and is close to unity only if the sub-graph edges' weights do not differ much, i.e., are internally coherent.

Shortest path length and diameter

The sequence of edges $(i, j), (j, k), ..., (p, q)$ defines a *path* between nodes v_i and v_q. The length of the path is just its number of edges. The *geodesic distance* (shortest path length) d_{ij} between two nodes is the length of the shortest possible path connecting them. The *diameter* d of a network is defined as the longest distance within it, $d = max[d_{ij}], \ \ i, j \in \mathbf{V}$. One has to be careful if the network has several components (groups of nodes without a connecting path between the groups), and exclude node-pairs from different components from any path calculations.

The small-world effect. It is hard to talk about distances in networks without mentioning the *small-world effect (swe)*, known also by the ever so famous nickname, "six degrees of separation". In most real-world networks, most nodes are connected by a short path, the *swe*. A fascinating experiment by Stanley Milgram in the 1960's showed that a letter given to a random person would take in average only six steps to reach another target individual. This shows up because single random edges can connect large groups of nodes which would otherwise be very distant in the network. See studies by, e.g., Watts and Strogatz [6].

Watts and Strogatz also gave the *swe* a more precise definition: it is apparent in networks where the mean geodesic length l scales logarithmically or slower with network size for a fixed mean degree. This is mathematically obvious, if for some center node the number of nodes inside distance r increases exponentially with r. This is true for many networks, for example for the Erdös - Rényi random graph model [8]. In the E-R model, the number of nodes is fixed, and each node-pair is connected with a constant probability p. Some networks are even shown to increase l no faster than $\log n / \log(\log n)$. An interesting study done recently with mobile phone user data shows that the probability of two users being connected decreases with the inverse square of the physical distance between them [9].

The *swe* combined with percolation theory might be one of the most interesting parts of complex network research for practical purposes. Take for example, the spreading of information or the spreading of a disease in a population. Naturally something spreads quicker the fewer steps it has to take. This analogy can be studied in a variety of ways, e.g. computer virus [10].

Network resilience. Percolation brings us next to network resilience, which is the effect on the network by removal of its nodes and/or edges. Most networks rely on the connections between the nodes to function, and with the removal of nodes the path lengths in the network will increase. Typically, at some point the paths will be so long that communication will break down. It may seem intuitive, but studies of Internet connections and web page networks show that if you remove nodes randomly it will take

a long time before it makes an impact on the mean geodesic length. If you target the nodes after decreasing degree on the other hand, only a small fraction need to be removed before the network collapses. This is particularly the case for scale-free networks (power-law degree distribution) [11]. The World Wide Web and the Internet are both scale-free networks, and the studies showed that removal of the 2% highest degree nodes resulted in a tripling of the mean geodesic length [12]. This has great relevance in epidemiology, where e.g., you want to target vaccination against a disease so that it gives the best effect. By vaccinating the highest degree nodes one will remove the most possible edges between carriers and not yet infected individuals. So how to vaccinate a population in the most effective way? We refer to the paper for the details, but vaccinating random people and then vaccinating (totally) random acquaintances of these, will eventually lead you to the highest degree nodes [13].

2.2 Communities

Many studies show that most social networks have a rich community structure, i.e., groups of nodes with a high density of edges between themselves, with lower density of edges to other groups of nodes. This definition is a bit vague, but it is common sense that people are more likely to interact with people of similar age, interests, occupations and so forth. This concept can be adopted by any type of network with different criteria, and since the criteria differs, there is no universal definition of a community. We operate with three scales in networks, the microscopic scale: nodes edges and their immediate surroundings, e.g. clustering, the macroscopic scale: features at the network level, e.g. degree distribution, hubs (very high degree nodes) and other global attributes, and the mesoscopic scale: structures which are dense with nodes and edges - communities, the backbone of the network. Communities are important because they often relate to functional units of a system. Examples includes groups of people interacting with each other in a society [14, 15], WWW pages related to similar topics [16], or proteins related to cancer metastasis [17].

The field of study and the applications of communities is still wide open and no universal definition of what constitutes a community has been adopted [18]. When you think of a social network, people tend to be part of different communities. The family can be one, the workplace another, and different groups of friends can be communities without any interconnection except one single person. All these communities are then *overlapping* at this node. Detecting overlapping communities requires more care from the algorithm.

Different methods and algorithms have been published for community detection. *Global methods* utilise the whole network structure for defining the communities, often optimizing a function depending on node configuration to come to the desired community structure. There are a variety of ways this can be done. *Local methods* take only the local topology[1]

[1]The structure and topology of a network is dependent on the edge configuration between nodes, path lengths, degree of nodes and other properties of the nodes and edges.

into account. Both types of method have their advantages and shortcomings.

2.2.1 Clique percolation

The algorithm used in this work for detecting the communities is called sequential clique percolation [19]. This is a development made from the original k-clique percolation method published by Palla et al. in 2005 [20]. A k-clique is a set of k nodes, all connected to each other. For example, a triangle is a 3-clique and in figure 2.1 we show cliques of different sizes. As discussed earlier, under the section *clustering coefficient and assortativity*, triangles are abundant in most networks, and it is easy to believe that there should be an even higher density of triangles within a community. The definition of a k-clique community is therefore the nodes which are in adjacent k-cliques. Adjacent means here that the nodes can be reached by the "rolling" of a k-clique, all the time containing $k-1$ of the previous nodes. See figure 2.2 for an illustration of such clique rolling.

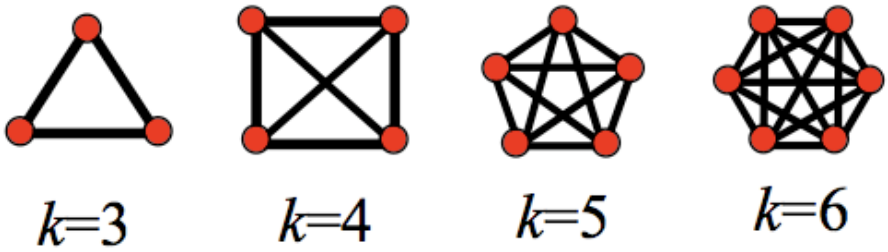


Figure (2.1): Cliques of different sizes. A k-clique is per definition a set of k nodes which are fully connected to each other. The figure is from the complex network course held by Jari Saramäki at TKK.

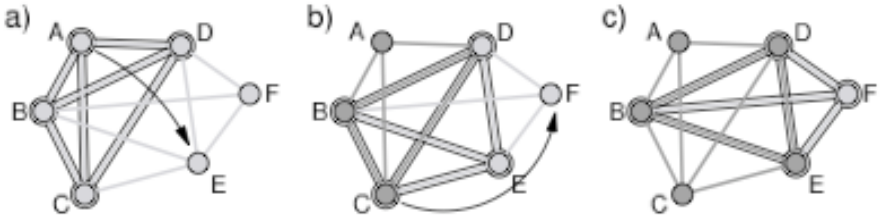


Figure (2.2): Illustration of the idea of k-clique percolation for $k = 4$. We see that $ABCD$ is a 4-clique. Releasing node A, we make a new 4-clique with node E, $BCDE$, and finally $BDEF$ by releasing node C. The new clique always contains $k-1$ nodes from the previous clique. All nodes which can be reached by this type of rolling belong to the same k-clique community. In this case we have a 4-clique community of size 6. The figure is included in [18], but the original source could not be determined.

The k-value is used to determine the community structure, triangles will usually give a structure with some very loose and large communities, while a higher k gives smaller,

but tight and dense communities. A k between 3 and 6 is usually preferred. One problem with this algorithm is that one can not detect hierarchical community structures, i.e., communities within communities. Running the algorithm again with higher k-value is how we will solve this, though. Note that communities can overlap each other with this method. Since it is based on topology only, the algorithm is deterministic and gives out the same result at each run. The result is easy to interpret and the resolution is easy to vary, with the k-value and the possible *thresholding* of edge weights (explained below).

The sequential clique percolation algorithm works by removing all the edges from the network and then putting them back one by one and updating the chosen k-community structure as it emerges. We refer to [19] for the details, but this algorithm is very computationally efficient compared to the original clique percolation algorithm which searches out the highest k in the network every time. By ordering the edges by weight one can also in one single run obtain the community structure for any threshold wanted, i.e., when only cliques with edges with weights above or below a certain threshold is added to the network. This is interesting, since e.g., it might not be reasonable that a few certain weak edges can hold big communities together. Full cliques can also be thresholded by intensity.

Chapter 3

The mobile phone data

The Department of Biomedical Engineering and Computational Science at TKK, has got some of the user and call/SMS statistics from a mobile phone subscription company in Europe. (From now on, the term *call* will be used as a common reference to both calls and SMS unless the context implies differently.) The different data covers different time periods from the first of January 2007. Naturally, since there is competition, this mobile phone subscription company does not have a 100 % market share. In this sence we have already lost some information, but we have to assume that such a large sub-group behaves more or less like the population as a whole.

Due to a no-disclosure agreement we are not at liberty to reveal any more information on the origin of the data, but we are free to use all the information we can learn from it. It is sufficient to know that this data is probably one of the best samples of this type we could have hoped to have in the last couple of years. On a note about privacy, we make clear that no content from the communication or actual phone numbers have been revealed to us. This makes it practically impossible to identify a single user from the data.

It is important to recognize that this data represents human behavior and that even though we can see clear trends in the data, some phenomena can not be easily explained in our field of science. This makes for a good opportunity for cooperation between network scientists and social scientists.

3.1 Description of the data

The original, unprocessed data consists of the following three parts:

3.1.1 Aggregated network

This data runs over a time period from the 1st of January to the 6th of May 2007, exactly 18 weeks, and is the sum of the number of calls and duration made from one user to another:

- Caller ID, which is unique and corresponds to a single phone number
- Receiver ID
- Number of calls in the period
- Number of SMS in the period
- Total length of calls in the period

How to transfer the data over to a network model should be fairly intuitive: each user is a node, and calls between users are edges. Here we have a clear directed network, because each call has a caller and a receiver. In the whole data material we have different information we can use as weight on the edges, the number and the duration of calls being the most obvious.

In the original, unprocessed data there are a bit over 6.5 million users. However, we want to make it a more "true" social network, where random interactions are ruled out. We do this by looking at the directed network, and remove all edges between nodes which point in only one direction, i.e, we remove the connections between users which have not been returned. We then remove all the users from the data which is no longer connected to the network (or never was in the first place, by not making or receiving any calls in the time period) and is left with a network which we now let be undirected. We can then choose a measure for the weights, and let the total between the users be the weights of the edges between them. This whole cleanup operation has led to a data set which is called the *mutual data* (calls has been mutual, i.e., made in both directions), and all our following analysis is based on this mutual data.

In the end we have 5 343 749 nodes in our network. Do note, though, that all the nodes are not connected in one giant component. There are many groups of users who have just been calling within themselves.

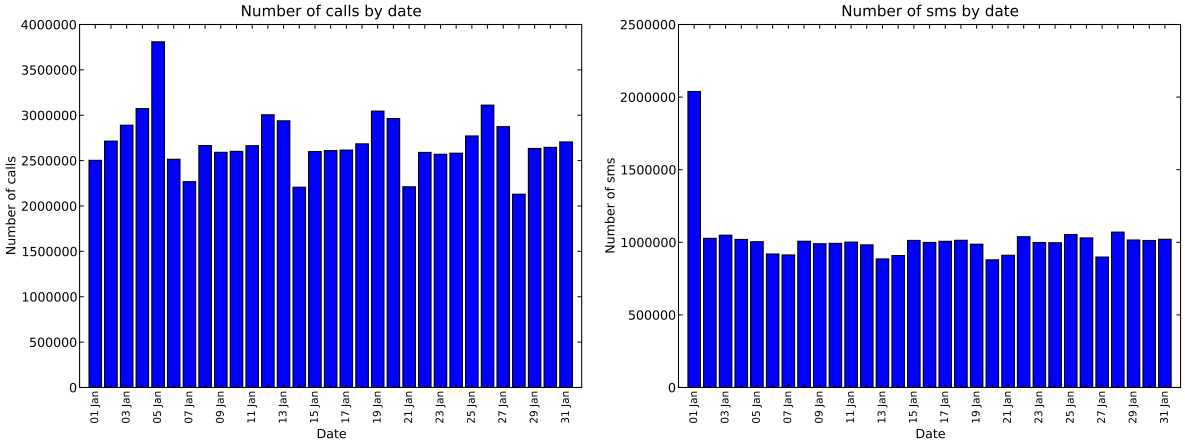
3.1.2 Event data

This data covers only the 31 days of January 2007, compared to the 18 weeks of the aggregated data, and describes each individual call made in this period. From this data we can see some of the calling patterns in the network. This data is also mutual and is presented in figure 3.1 and 3.2:

- Event date
- Event time
- Caller ID
- Receiver ID
- Type, call or SMS

- Duration, 0 for SMS

The total number of calls for the whole 18 week period is roughly 384 million, which leaves us about 64 million calls with detailed information. For SMS the numbers are, respectively, 133 and 22 million.



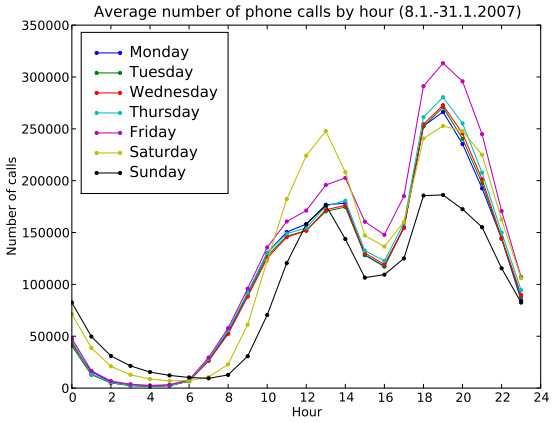
(a) Calls (b) SMS

Figure (3.1): The number of calls made and SMS sent on each day of January 2007. (a) The first day is a Monday, the sixth day a public holiday (explaining the peak at the fifth), the spike pairs are Fridays and Saturdays and the muted spikes are Sundays. (b) Note the number of SMS sent on new year. The muted spike pairs are Saturdays and Sundays, but the reason for the sudden increase in SMS on Sunday the 28th is not known. We see that both plots have a deviating shape during the first week of January, which is maybe not so surprising.

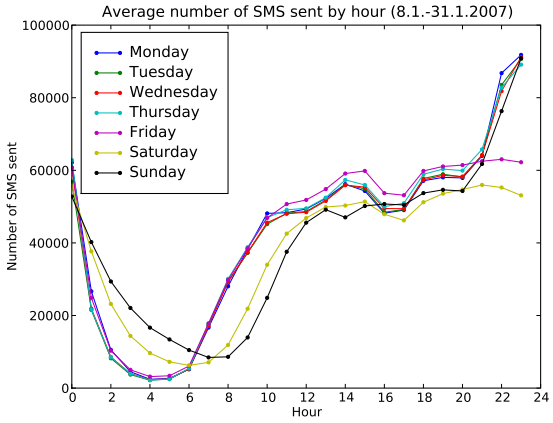
3.1.3 User demographics

In theory, for each mobile phone user in the data we should know the following:

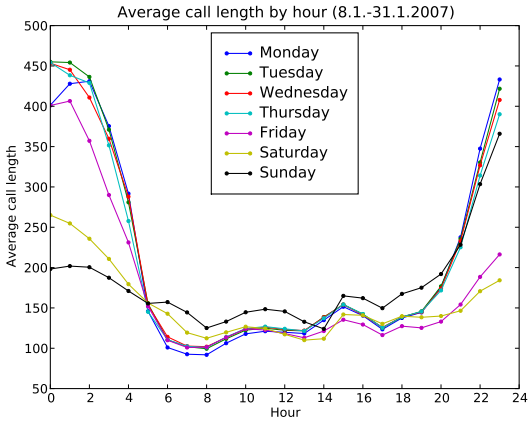
- ID, as mentioned earlier
- Age
- Sex
- Zip code
- User subscription type, pre or postpaid



(a) Number of calls made



(b) Number of SMS sent



(c) Call length

Figure (3.2): The average hourly statistics for each weekday from the 8th to the 31st of January. The first week of January has been filtered out. As seen here and in figure 3.1, calls and SMS play different roles in mobile phone communication, naturally (but maybe non-trivially) related to the social implications of their technical differences.

Some age, sex and prescription statistics are shown in figure 3.3. There is some incomplete information for a number of users in form of not valid zip codes, age zero years and unknown sex. This is probably mainly due to the nature of sloppy registration of people who buy a prepaid subscription. From our data, the default entry for a prepaid user seem to be: a man of age zero without a valid zip code. As there are roughly 1.3 million such prepaid users without known information, it is clear that they increase the uncertainty of any analysis incorporating the demographic information. Phone numbers may also change between actual persons when people end their subscriptions, but the number of such occurrences is assumed to be so low that it can be disregarded as noise. All in

all, there are about 2.2 million men and 1.8 million women with complete demographic information out of a total of 5.3 million people in our network.

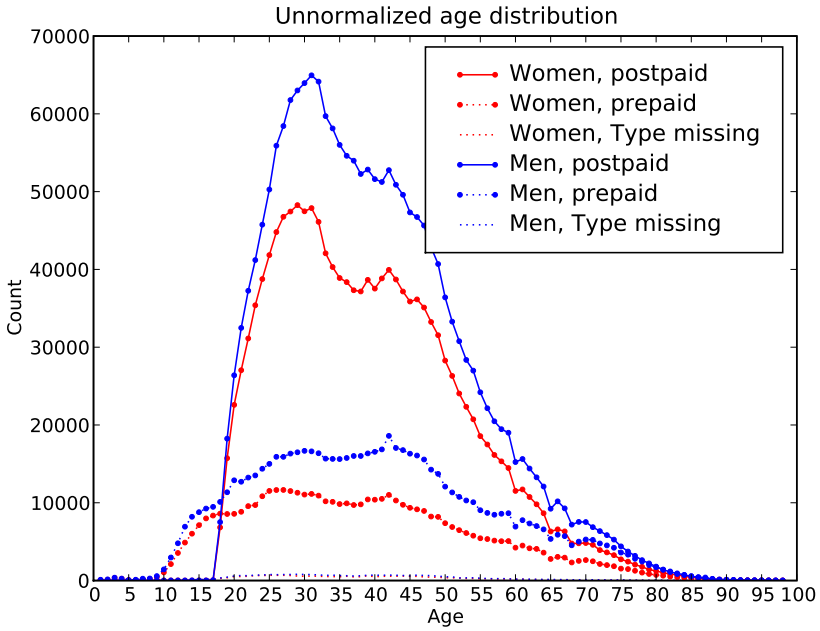
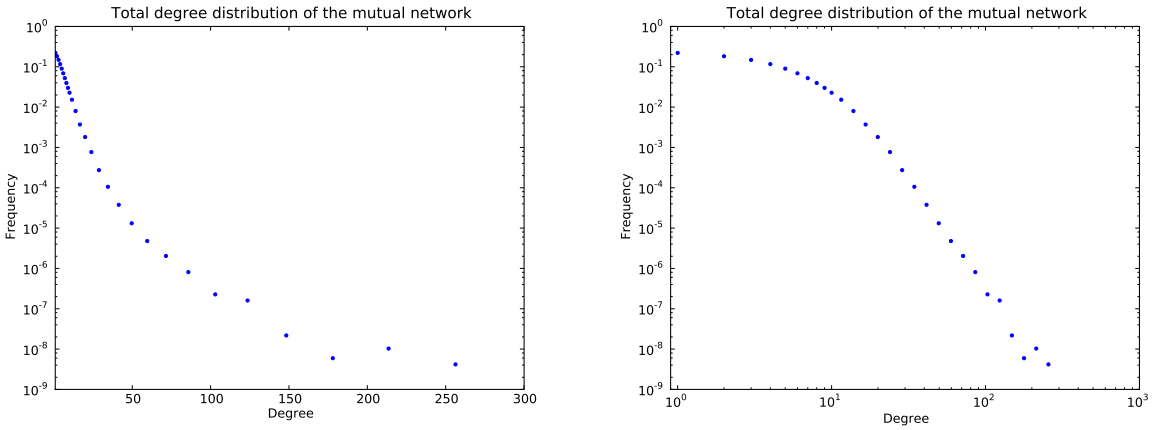


Figure (3.3): The number of users in the real network at each age and their sex. This is before the network is made mutual, but should reflect the general distribution of the mutual network. Prepaid users with age zero and unknown sex have been cut from the plot.

3.2 Basic analysis

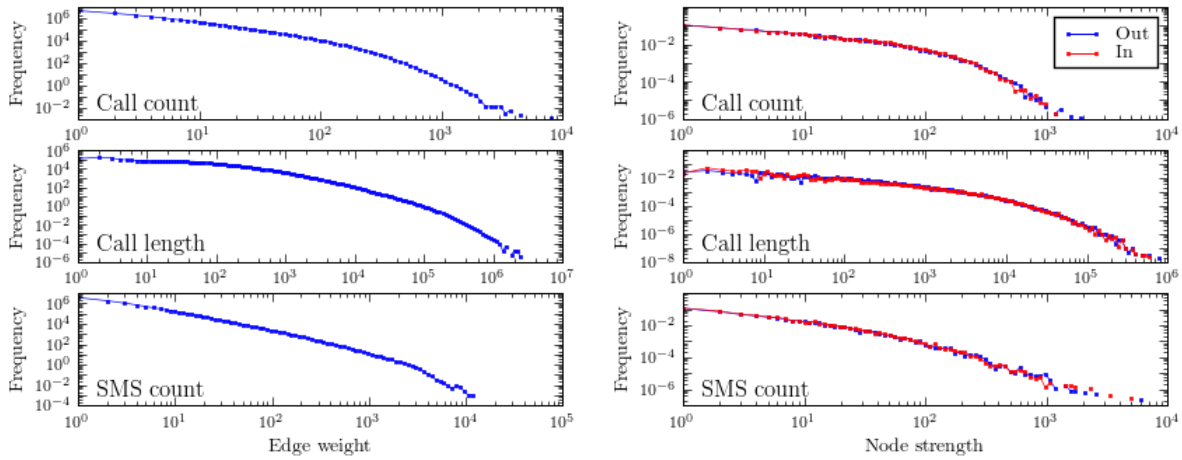
In figure 3.4 and figure 3.5 we present the first and elementary steps of the analysis of the mobile phone network.



(a) Degree distribution, linlog plot

(b) Degree distribution, loglog plot

Figure (3.4): Degree distribution of the aggregated, mutual phone network. (a) The degree distribution has almost exponential behaviour for degree up to 20, corresponding to an exponential distribution $p(k) \sim e^{-0.25k}$. (b) Typically for a social network, we can see a long tail of nodes with degrees well above the average, $\bar{k} = 4.476$. This corresponds to a power-law $p(k) \sim k^{-5.5}$. What these figures not so clearly tell is that only 0.6 % of the nodes have a degree above 20, making the majority of the node degrees exponentially distributed.



(a) Weight distribution

(b) Strength distribution

Figure (3.5): Weight and strength distributions of the aggravated network. (a) Three possible weights is used for calculating the strength distributions (sum of the weights of the edges of a node). (b) We have both incoming and outgoing strength. The strength averages over the 18 weeks are: call count - 65.6, call length - 9741 s or 163 m, SMS count - 23.8. Again the fat tailed distributions give rise to certain conclusions, for example with call counts, 69 % of all users make less than the average number of calls, and 0.29 % of users make more than 10 times the average number of calls.

Chapter 4

Visualization of communities in the phone call network

Since communities is such an important part of any social network, we have decided to do an extensive visual study of communities in our network. We have discussed in section 2.2 what a community is and everyone can recognize social communities in daily life. But how do they look in the actual network topology when we deal with networks in theory? Having a good visual background will help us in the further research on communities.

Our phone call network contains over 5.3 million nodes with almost 12 million edges. Since it is so large and also consists of many disconnected components, it is not feasible to present a plot of the whole network. We have therefore concentrated our visualization on the four largest 4-clique communities found by the clique percolation algorithm. These communities have sizes from 229 to 518 nodes. We have chosen $k = 4$ as the lowest k-value to study. When we tried $k = 3$, we found that more than 1.5 million people were in the largest community. Naturally, that does not rhyme with reality.

Just plotting the 4-clique communities, saying that they are large and be done with it, would be pretty dull. It is expected, however, that such big communities would show some kind of nested community structure, or hierarchical community structure. Common sense says that within large groups of people who have something in common, there are smaller groups of people who just know each other better. By running the clique percolation algorithm on these communities again for cliques of size 5 and higher, we see that a remarkable inner community structure reveals itself in the plots.

4.1 Himmeli graph plotting software

Automatic graph drawing is one of the advantages of computers in our field. The software used to make the plots in this work is Himmeli[1] created by V-P. Mäkinen from TKK. Himmeli is based on a force-directed algorithm which is capable of making an acceptable

[1]Himmeli v3.3 for Linux could as of May 2009 be found on the web page: http://www.finndiane.fi/software/himmeli/

2D node layout for almost any type of network. The input to Himmeli is a configuration file where all your parameters and instructions to the program are specified. The most important parameter is of course the edge-file[2] of the network. Other parameters let you influence the visuals of the output, or let you do more advanced modifications like thresholding of edge weights (which we will show). Since the algorithm in Himmeli for creating the optimal layout of the nodes is time dependent and generally not deterministic, we have used the output of a single run to fix the coordinates of the nodes in the next plots. This means that all the plots of the same 4-clique community have the nodes in the same places.

4.2 Description of the plotting scheme

For each of the four 4-clique communities we will present four plots. Note that not only the edges needed to create the actual k-clique community is plotted, but also other edges within the community which not necessarily make cliques. The first plot is just the generic plot of the 4-clique community, see e.g. figure 4.1. The weight used on the edges is the total number of calls and SMS between two users. We choose to use this weight as it is the most straightforward measure of the strength of the relationship between two users. We could also have used the duration of calls, but we would have had to give some relative weight to those who have only sent SMS. Right now we decided against trying this approach, as deciding the relative strength between calls and SMS is maybe a job better suited for sociologists. The color of the edges represent their weights, the higher the weight of an edge is, the darker its color will be. The thickness also increases slightly with weight. We have used the demographic data described in section 3.1.3 to give each node a look after the following system:

- White circles are users without reliable age, sex and zip code information.
- Squares are men, triangles are women, users with reliable information.
- The colors give the ages of the users: yellow: under 30, green: 30-39, blue: 40-49, red: 50 and up.
- The sizes of the nodes represent the degree (the number of edges attached to the node), the bigger the node, the higher the degree.

In the last three plots the different k-clique communities detected have been plotted on top of each other (see, e.g., figure 4.2, 4.3 and 4.4). One has to remember that a k-clique has to consist of several $k-1$ and lower ordered cliques, so in the plots the higher cliques' colors are just shadowing for the lower cliques' colors. Each k-clique community have

[2]The edge-file is the common way to store a network. It is all you need to use your favourite computer algorithms to make a representation and analysis of it. Usually, each line in the file describes an edge on the format: edge source (node n) edge destination (node n) weight (if any). The size of the edge-file of our phone call network, after the earlier mentioned modifications, is just over 200 MB.

been given its own color for the edges and the basic colors are the same in all four 4-clique communities:

- Yellow for the 4, green for the 5 and blue for the 6-clique communities. Other colors used for higher ordered clique communities are described in the relevant figure texts.

The coloring of the edges lets us see which nodes belong to the nested communities. If a node has edges to it from one color, it also belongs to the looser underlying communities of lower ordered k-cliques.

4.2.1 Thresholding the edge weights

In the two last figures of each 4-clique community we have removed a fraction of the weakest edges (lowest weights) to see how the communities hold together then (see, e.g., figure 4.3 and 4.4). First 40 % and then 70 % of the weakest edges have been removed.

The motivation for doing this goes all the way back to the hypothesis published by Granovetter in 1973: weak edges tend to contribute to a low number of triangles, while strong edges contribute to a high number of triangles [21]. This means that the proportional overlap of two persons' individual friendship networks varies directly with the strength of the edge between those two persons. Mobile phone user networks have later been studied extensively, and we will tie our work together with the roles of weak versus strong edges found in such networks in [22, 23].

By studying information flow and the fragmentation of the network by removal of edges, the mentioned studies conclude that weak edges tie communities together, while communities consist of stronger inner edges. The *overlap* of an edge e_{ij} is introduced as a measure of how many of their neighbours do i and j have in common. It is shown that strong links have higher overlap, which is natural to be found inside communities [22]. The next study uses the concept *betweenness centrality* for an edge, which is a measure of how many of the shortest paths between nodes in the network goes through the edge. Naturally, edges which are inside a local community have more or less the same low betweenness centrality, since there are more paths to choose from, while an edge connecting two separate communities has to connect all the nodes, and will have a marked higher betweenness centrality. It is shown in [23] that edges with high betweenness centrality are weaker links, and therefore connections between communities.

The question which remains is whether the strongest links within a community is associated with the even denser sub-communities, and we can visually confirm that they are.

4.3 The visual plots

To help the reader enjoy the visual plots to the fullest, we will present the information we have learned from them before the actual plots. This will hopefully ease the experience of the otherwise somewhat cryptic plots, and help the reader to know what to look for. To

avoid cluttering in the plots we decided against putting any kind of label on the nodes. We are aware of the problems this creates regarding referring to the figures, so rather than giving lengthy descriptions to where one can see this or that phenomenon, we just hope that the reader has enough time to let the visual plots sink in together with the following résumé.

4.3.1 What we have learned from the visual plots

First of all, we are extremely pleased with the quality of the plots. The hierarchical community structures are presented clear as day. This gives us a good feeling about the use of the clique percolation method for detecting communities in a large social network. Also, it is a true joy to be able to actually *see* the consequences of the Granovetter hypothesis in effect. The number of edges between the different communities is clearly reduced during the thresholding, confirming the earlier studies on the subject. There is no doubt that the amount of yellow in the plots is more reduced with the increased edge weight threshold than the other colors. The structure of the higher ordered communities are still clearly present, even at the 70 % threshold. The structure obviously breaks at higher thresholds, but the higher ordered communities keep more single edges, which we show in the numerical analysis in figure 4.17. This indicate that there is a core or cores in a community, which is a topic currently under study.

Note that the communities are not required to be perfect clique communities anymore when we do thresholding, we simply remove the weakest edges and replot the figure. It is just logical that any community will shrink bit by bit when the threshold is increasing. Following the strict definition of k-clique communities would quickly leave us without any communities at a very low threshold.

Overlapping communities are abundant in our plots, but they are quite hard to spot due to the huge number of crisscrossing edges. Two certain examples of overlapping communities have guides so one can spot them in the figure text of figure 4.10 and 4.14.

It takes only a quick glance to understand that the demographic part of our plots is worth close to nothing. There is an overweight of white circles in three of the four 4-clique communities. Other than that, there seems to be a lot of young people in these big communities and the only thing conclusive we can say about the demographic properties is that the valid zip codes are almost always within the same area. We have chosen to not plot the zip codes, but write a note about it for each community in the figure text. For further analysis of the demographic properties we resort to numerics in the next chapter.

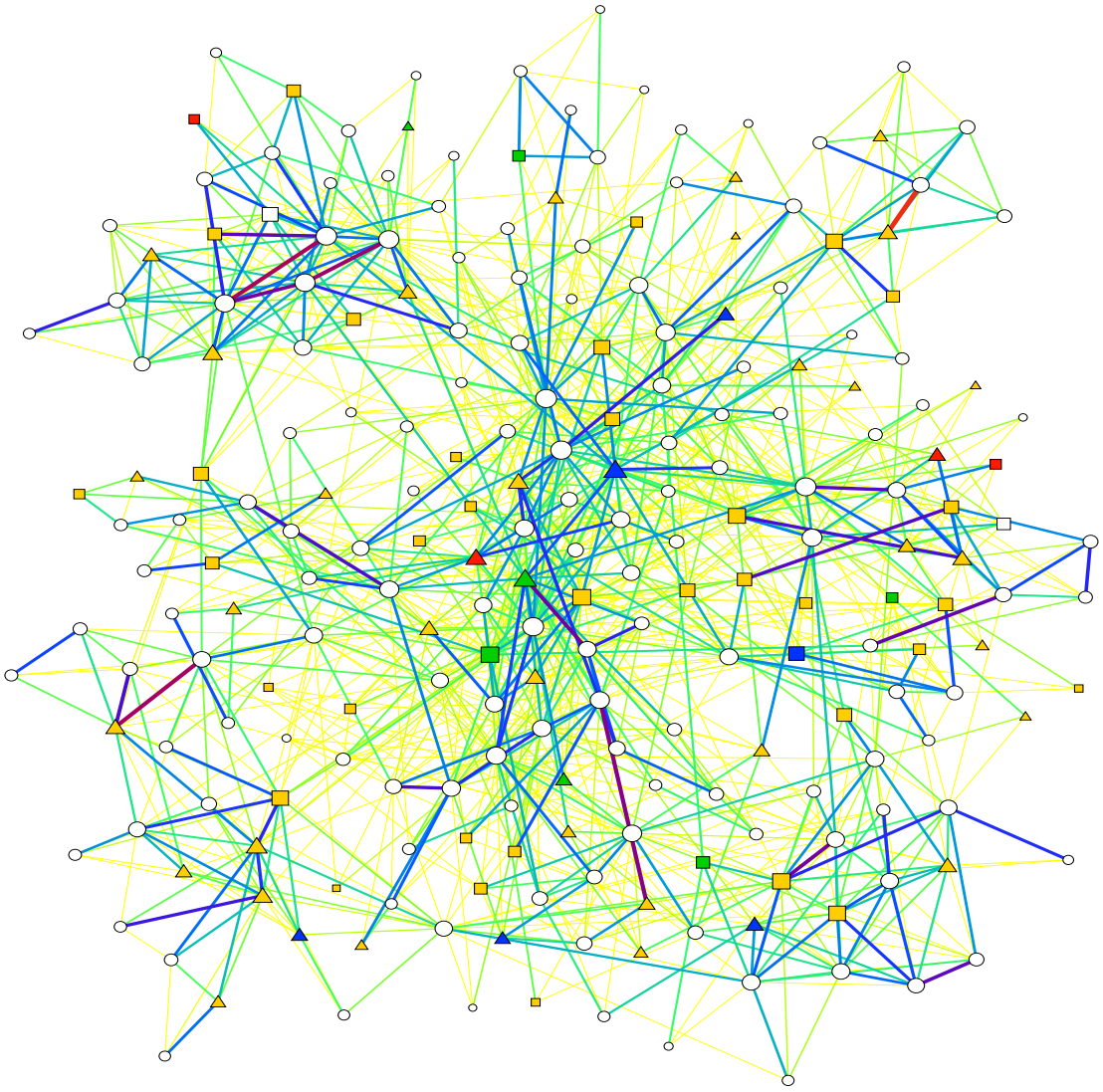


Figure (4.1): The fourth largest 4-clique community, 229 nodes, ∼1200 edges. Darker colors mean stronger edges. The color of the node gives age, in ascending order: yellow, green, blue, red, triangles are women, squares are men and white circles are users without information. Of the valid zip-codes, only a few were not within the same area.

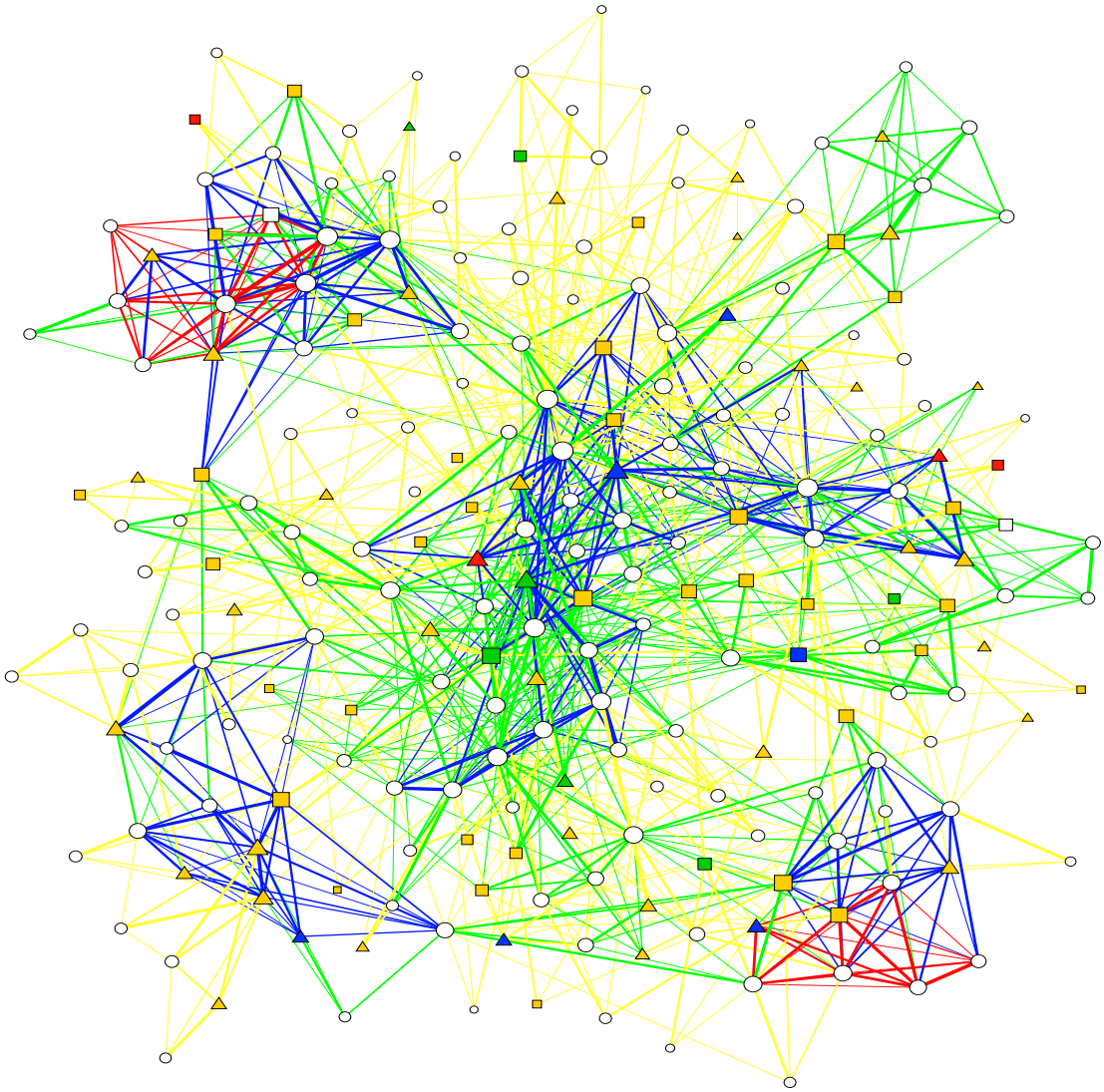


Figure (4.2): Plot of the 4-7-clique communities in the 4th largest 4-clique community. Red denotes the 7-clique communities.

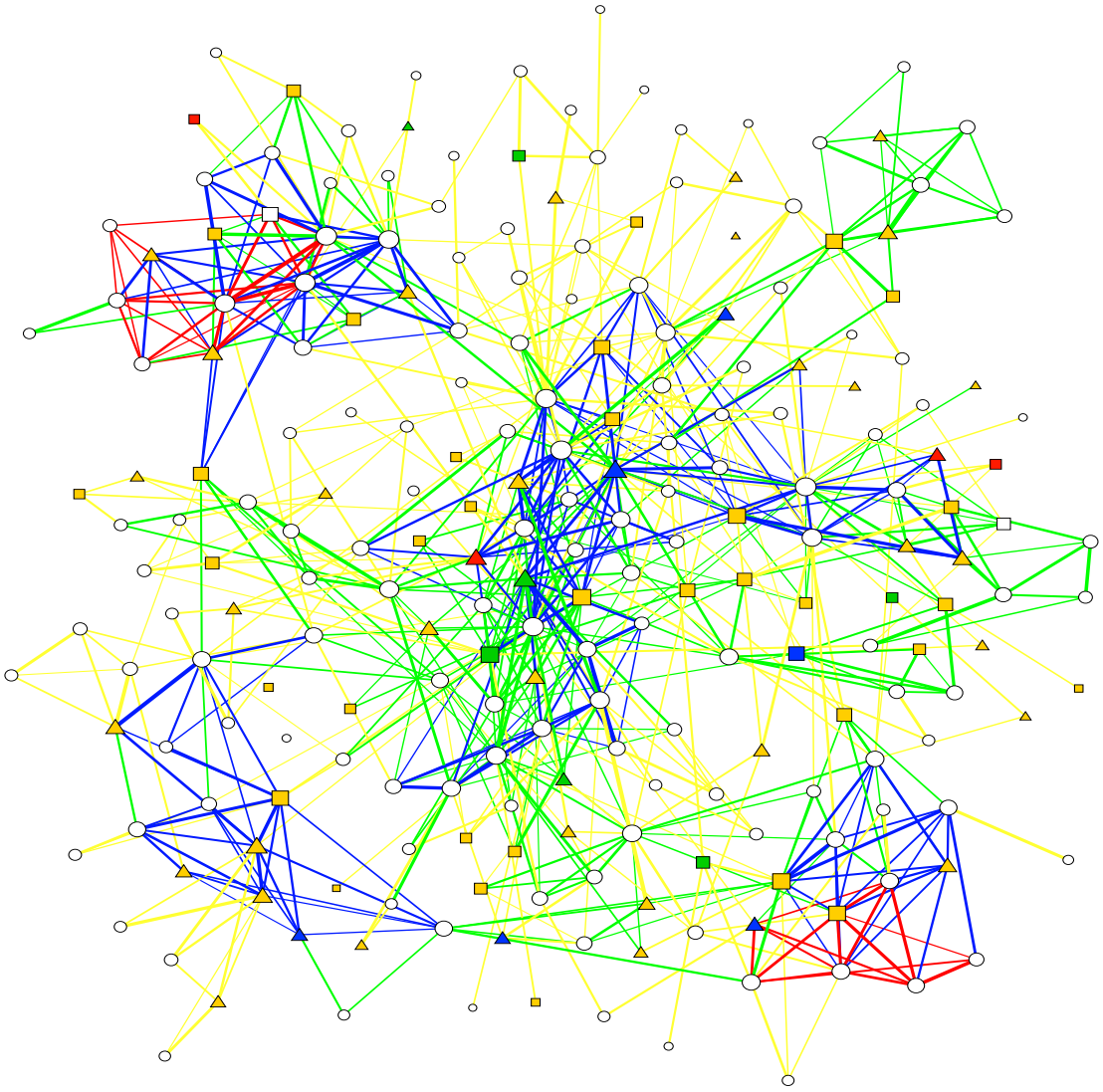


Figure (4.3): Plot of the 4-7-clique communities in the 4th largest 4-clique community. Red denotes the 7-clique communities. The 40 % weakest edges have been removed.

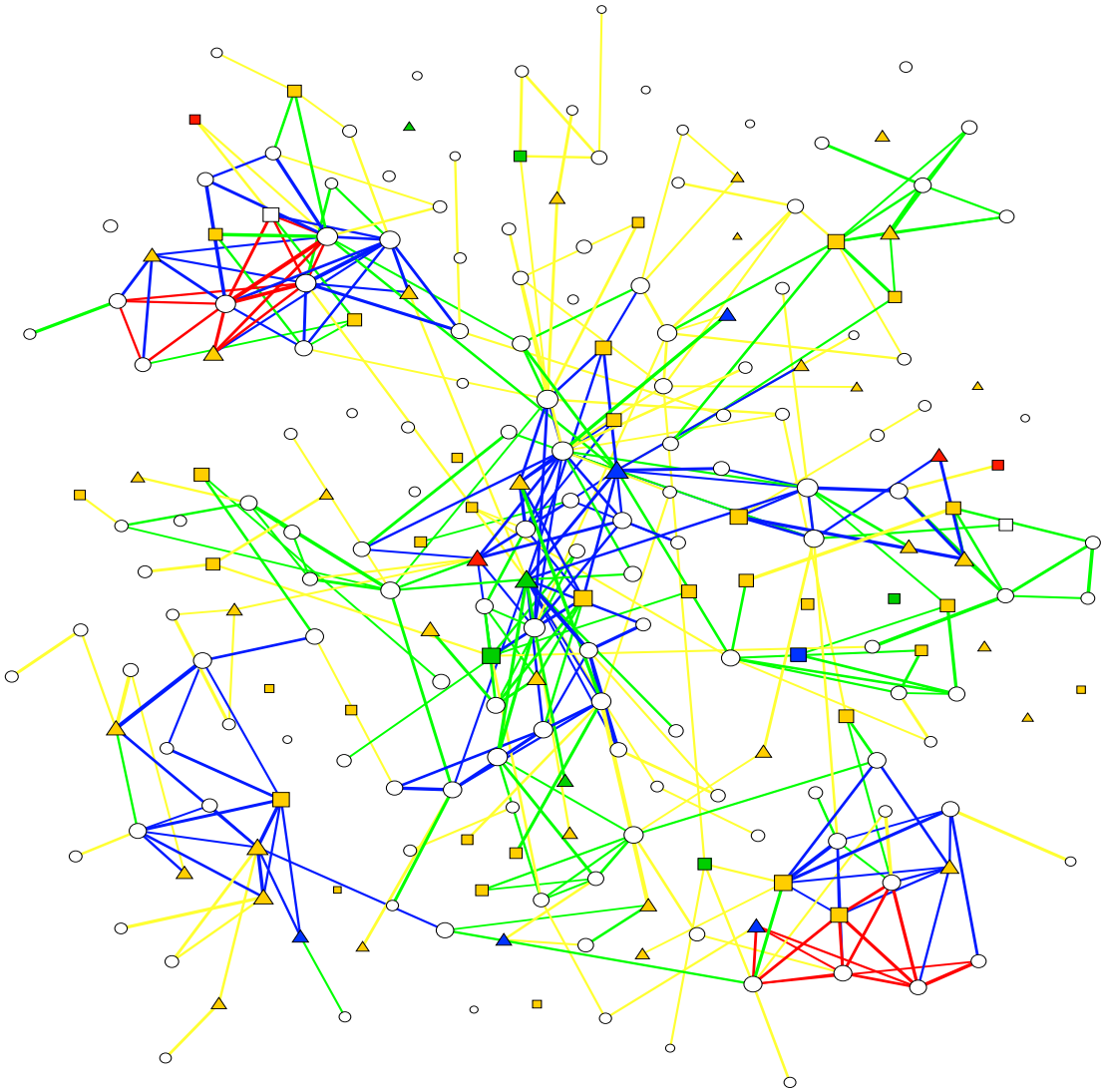


Figure (4.4): Plot of the 4-7-clique communities in the 4th largest 4-clique community. Red denotes the 7-clique communities. The 70 % weakest edges have been removed.

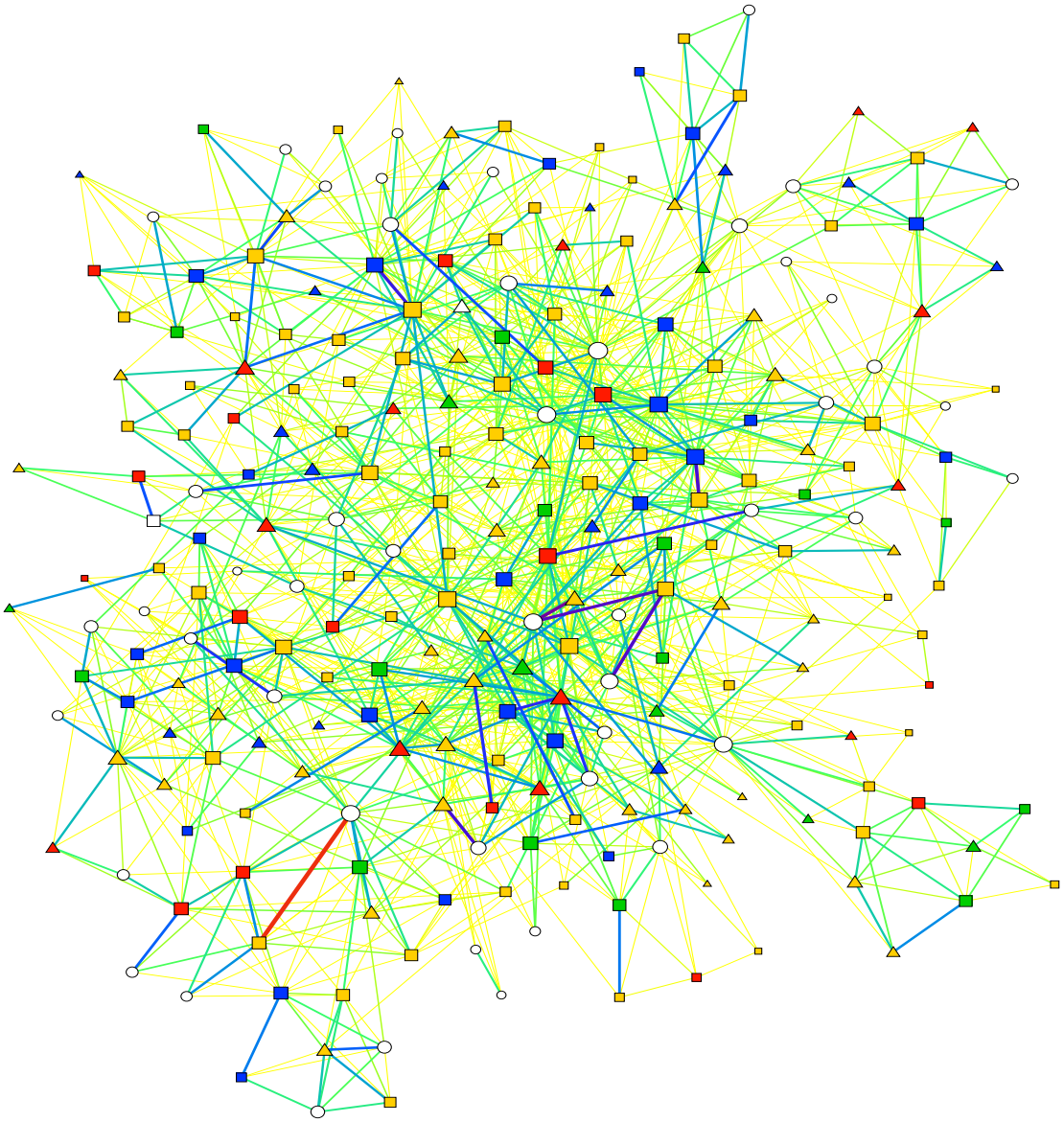


Figure (4.5): The third largest 4-clique community, 254 nodes, ∼1350 edges. Darker colors mean stronger edges. The color of the node gives age, in ascending order: yellow, green, blue, red, triangles are women, squares are men and white circles are users without information. Of the valid zip-codes, only a few were not within the same big city.

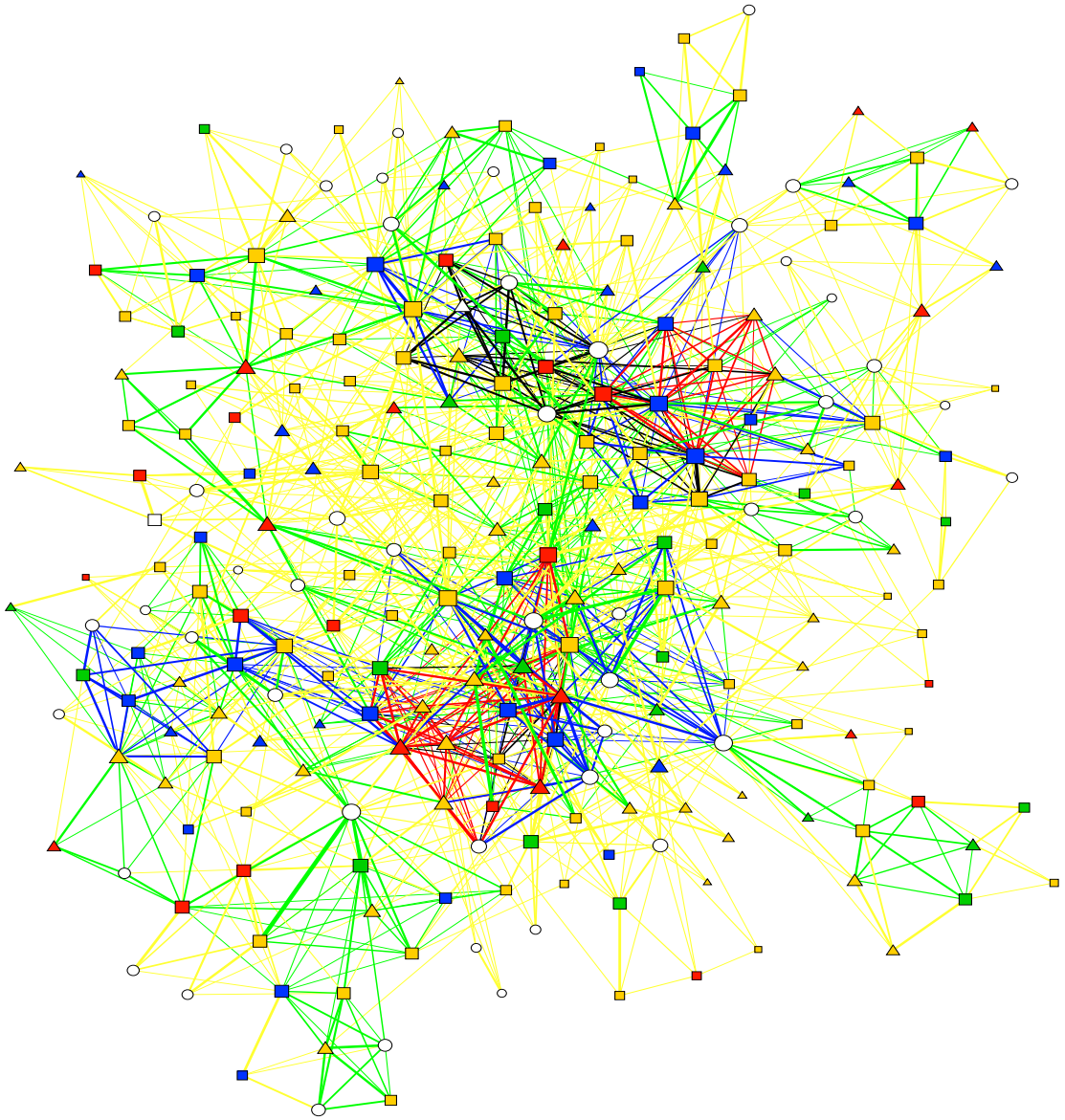


Figure (4.6): Plot of the 4-8-clique communities in the 3d largest 4-clique community. Black denotes the 7 and red denotes the 8-clique communities.

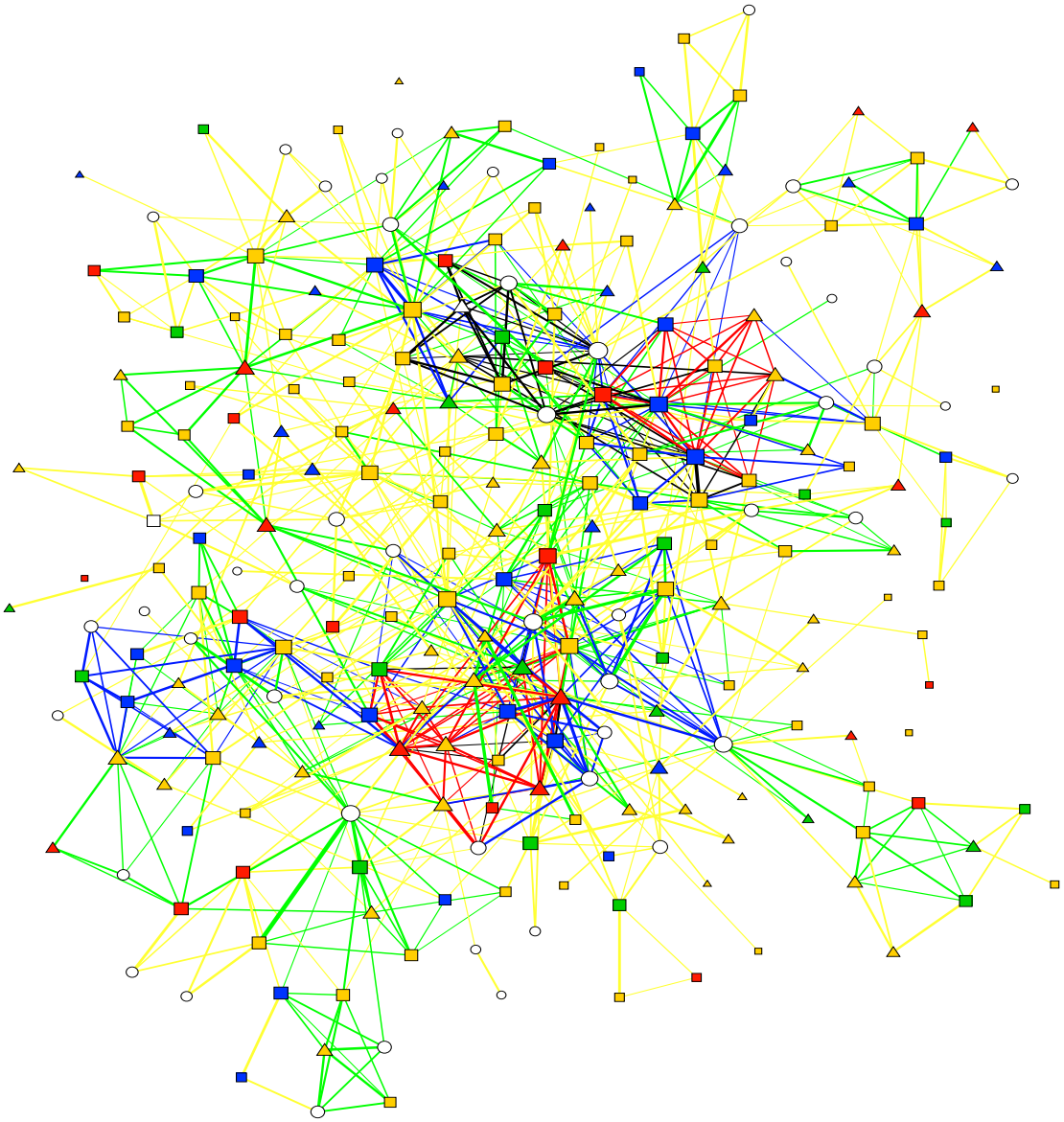


Figure (4.7): Plot of the 4-8-clique communities in the 3d largest 4-clique community. Black denotes the 7 and red denotes the 8-clique communities. The 40 % weakest edges have been removed.

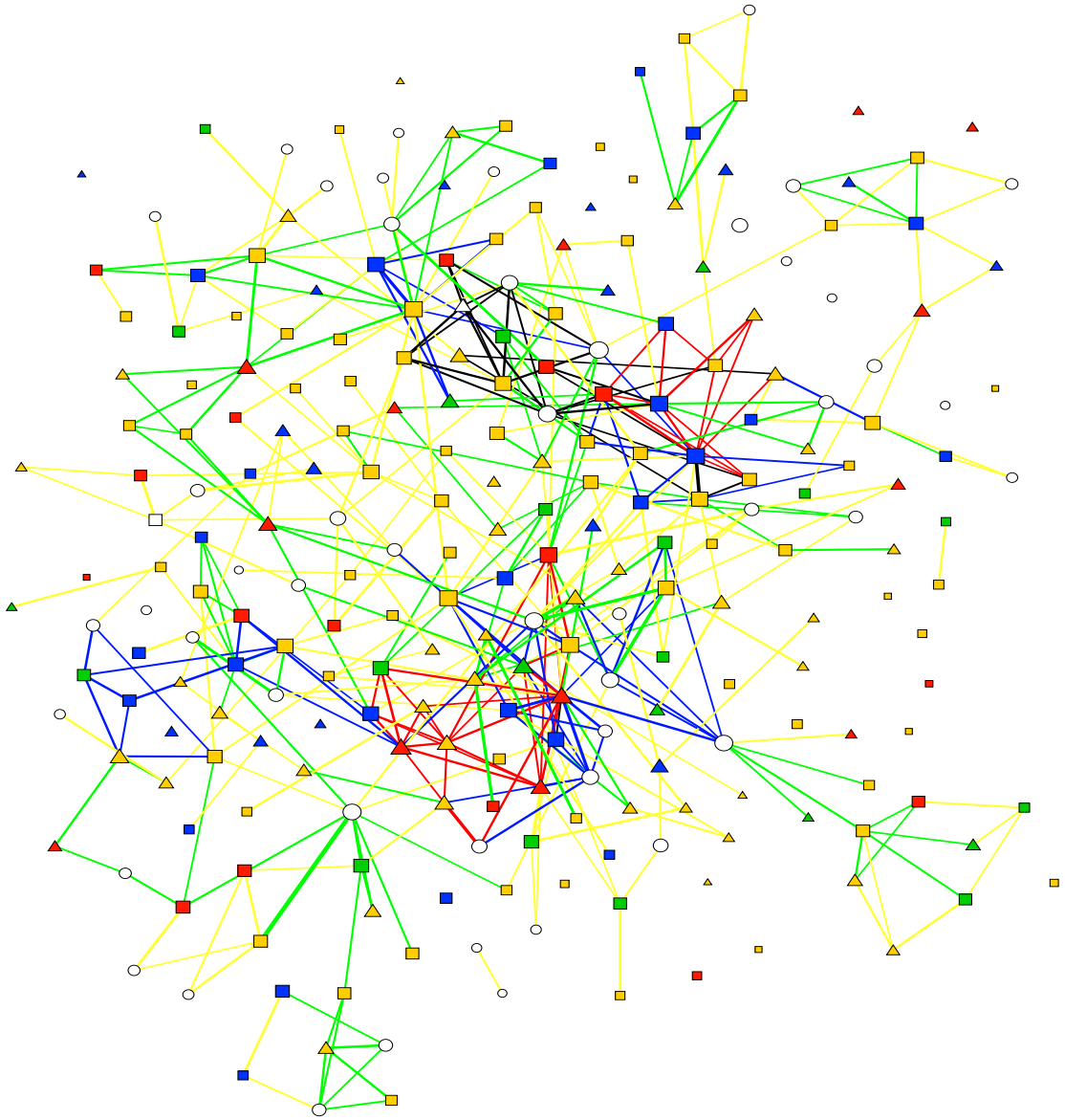


Figure (4.8): Plot of the 4-8-clique communities in the 3d largest 4-clique community. Black denotes the 7 and red denotes the 8-clique communities. The 70 % weakest edges have been removed.

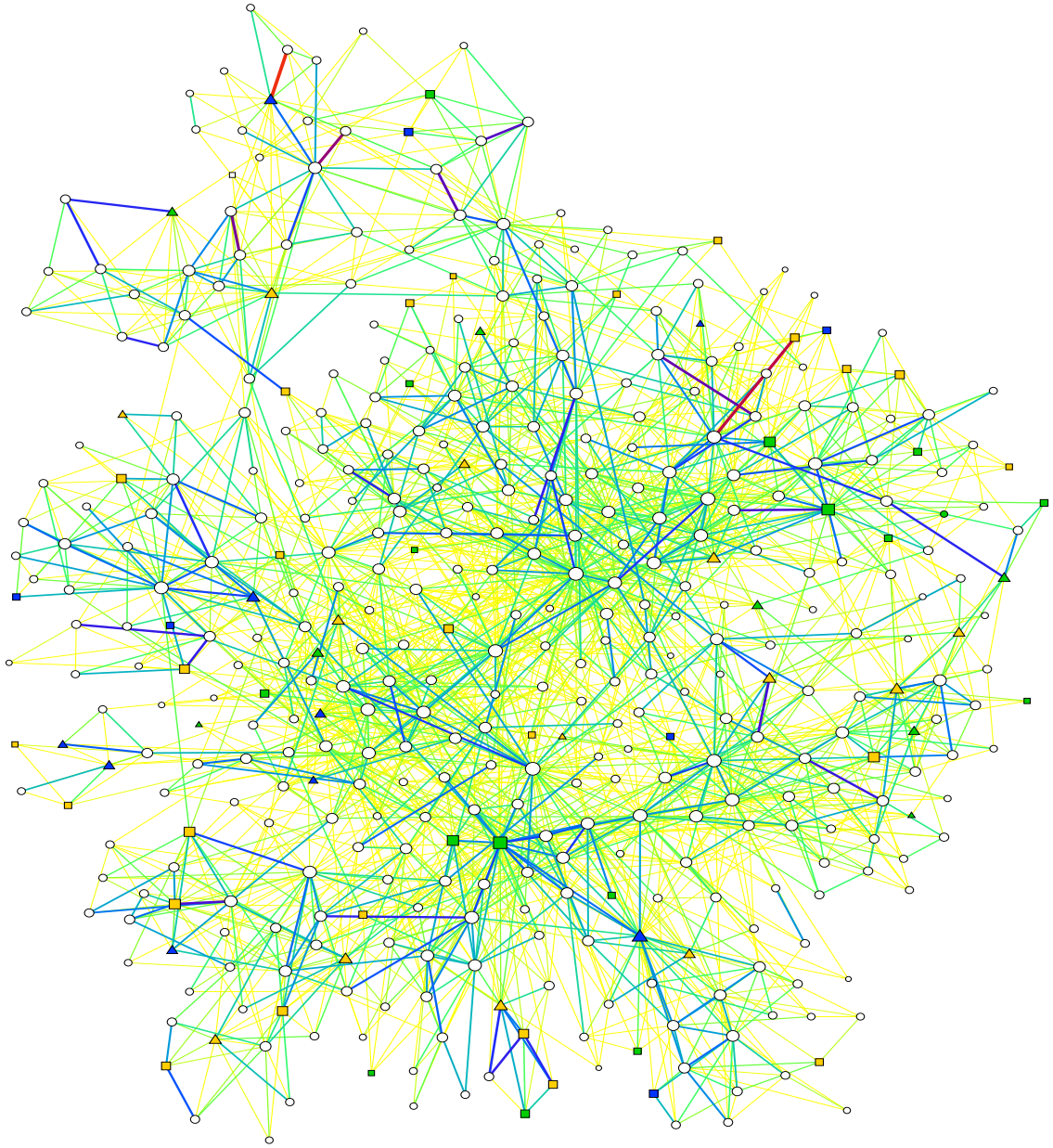


Figure (4.9): The second largest 4-clique community, 430 nodes, ∼2300 edges. Darker colors mean stronger edges. The valid zip-codes seem to fall into 5 distinct groups.

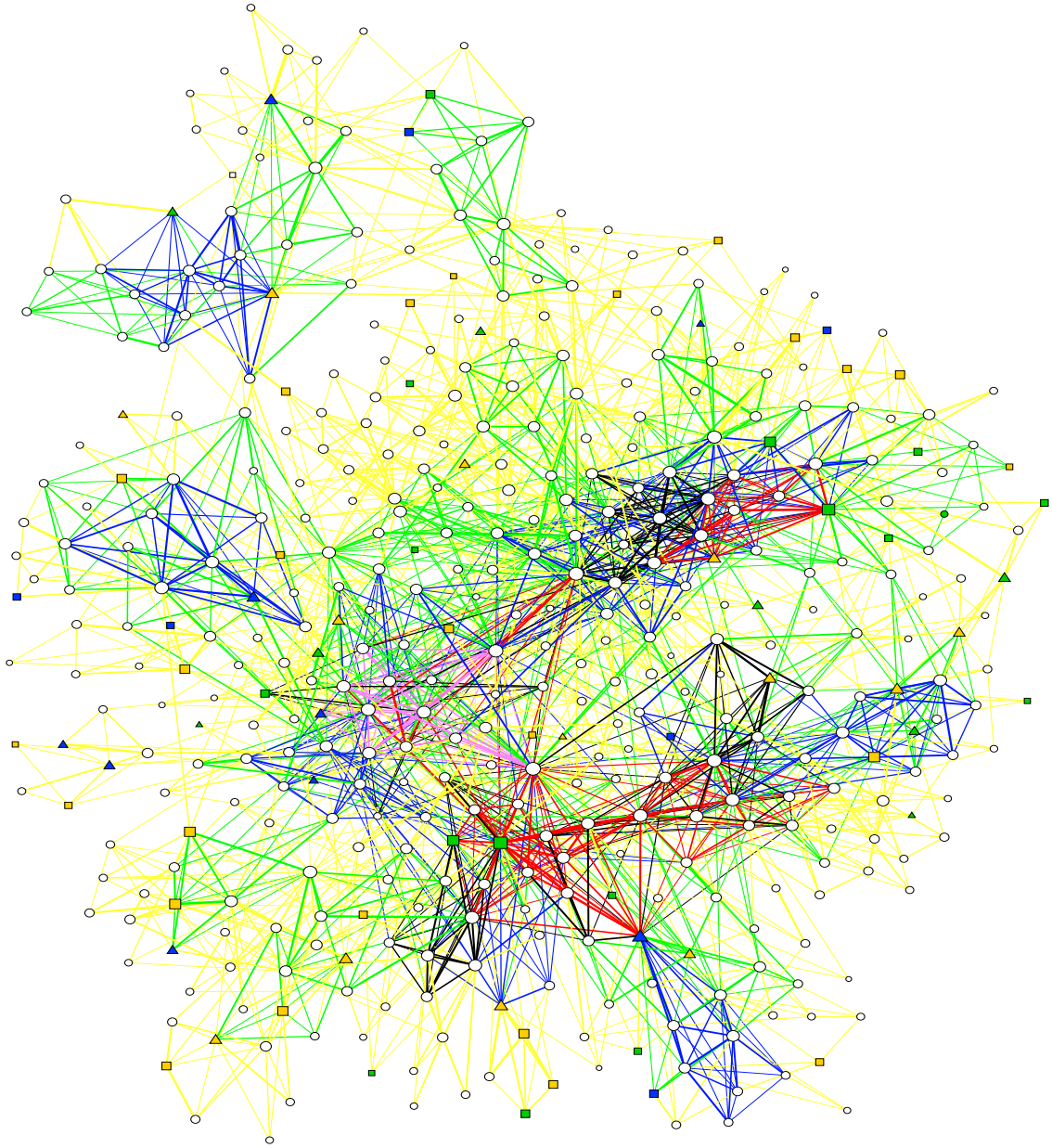


Figure (4.10): Plot of the 4-9-clique communities in the 2nd largest 4-clique community. Black denotes the 7, red the 8 and pink denotes the 9-clique communities. The green sub-net at the top in the middle (including one green, one blue square and seven white circles) consists of two overlapping communities. The five lower nodes are one community and the seven upper nodes are another.

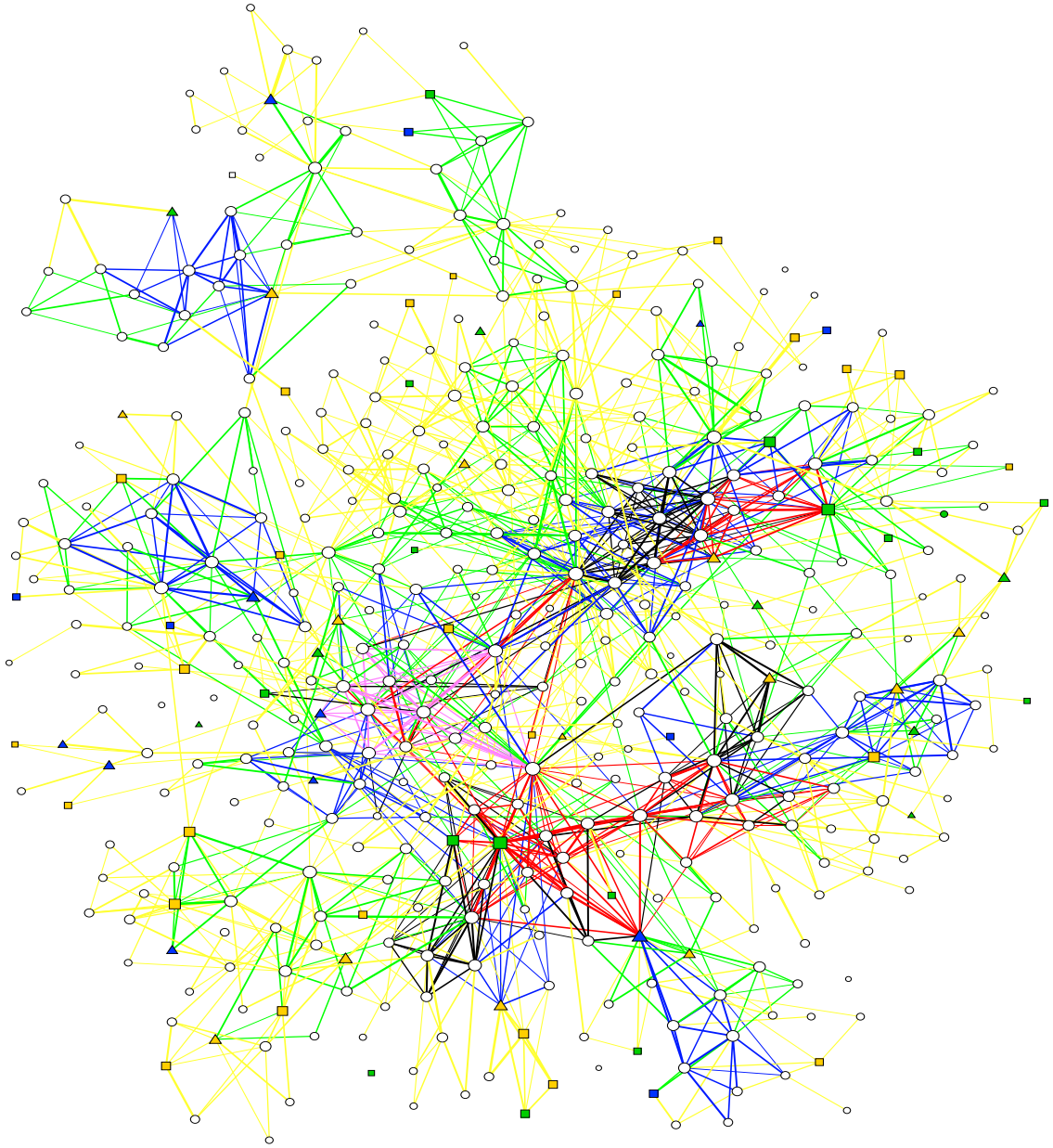


Figure (4.11): Plot of the 4-9-clique communities in the 2nd largest 4-clique community. Black denotes the 7, red the 8 and pink denotes the 9-clique communities. The 40 % weakest edges have been removed.

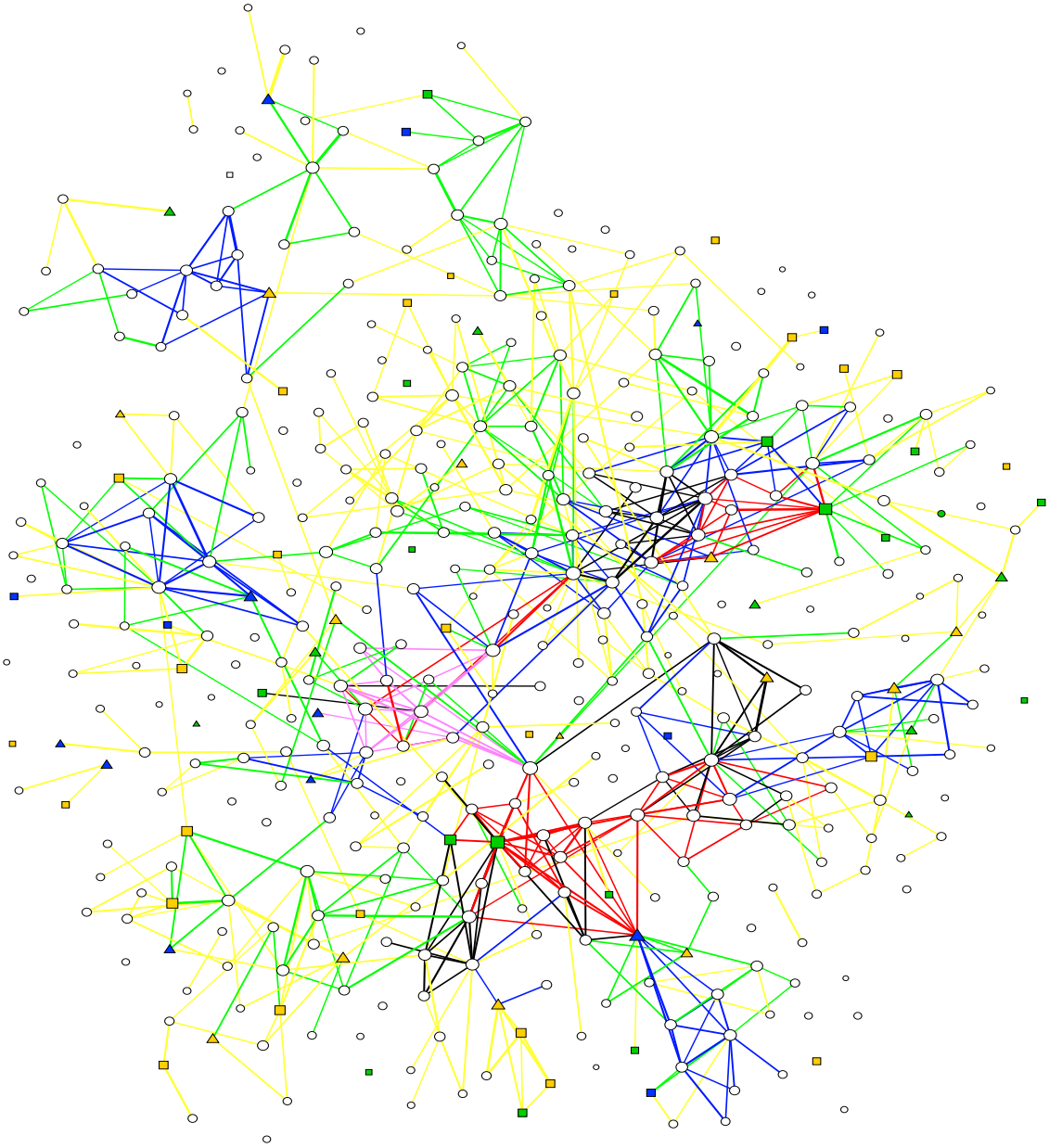


Figure (4.12): Plot of the 4-9-clique communities in the 2nd largest 4-clique community. Black denotes the 7, red the 8 and pink denotes the 9-clique communities. The 70 % weakest edges have been removed.

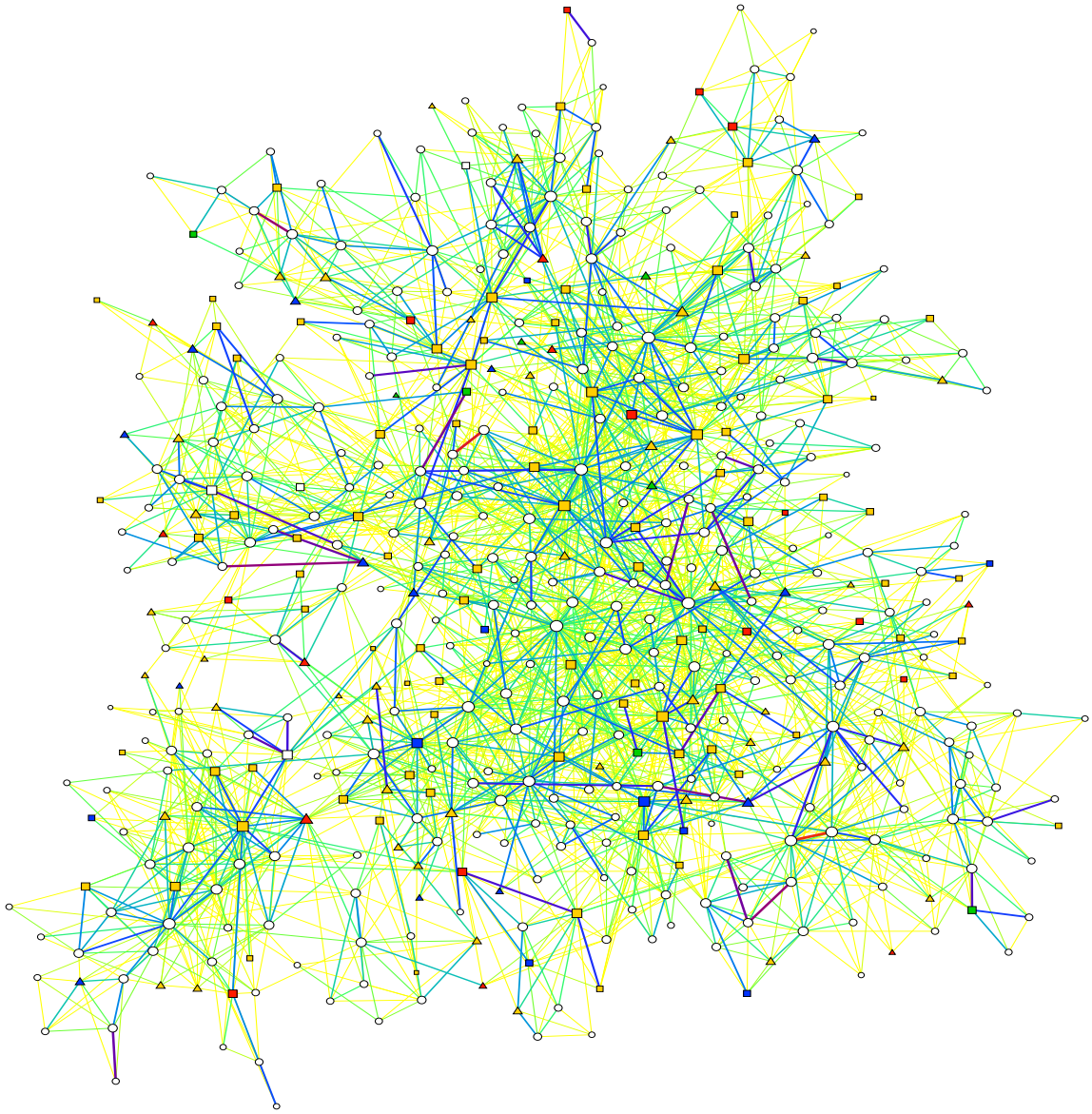


Figure (4.13): The largest 4-clique community, 518 nodes, ∼2800 edges. Darker colors mean stronger edges. Of the valid zip-codes, only a few were not within the same big city.

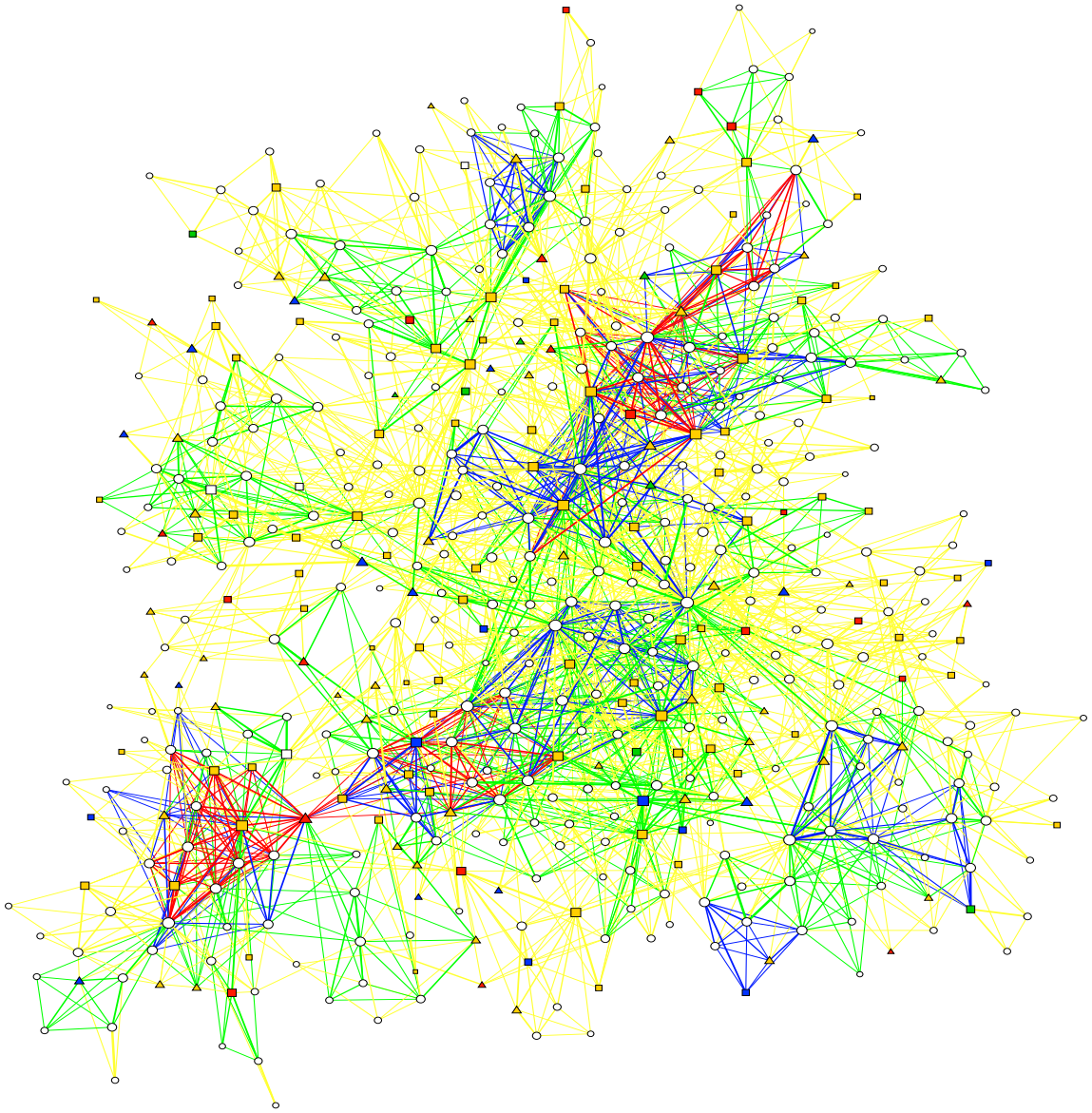


Figure (4.14): Plot of 4-7-clique communities in the largest 4-clique community. Red denotes the 7-clique communities. Note the red triangle node between the two lower 7-clique communities. This person belongs to two clearly distinctive 7-clique communities. This is probably the best example of overlapping communities in our plots.

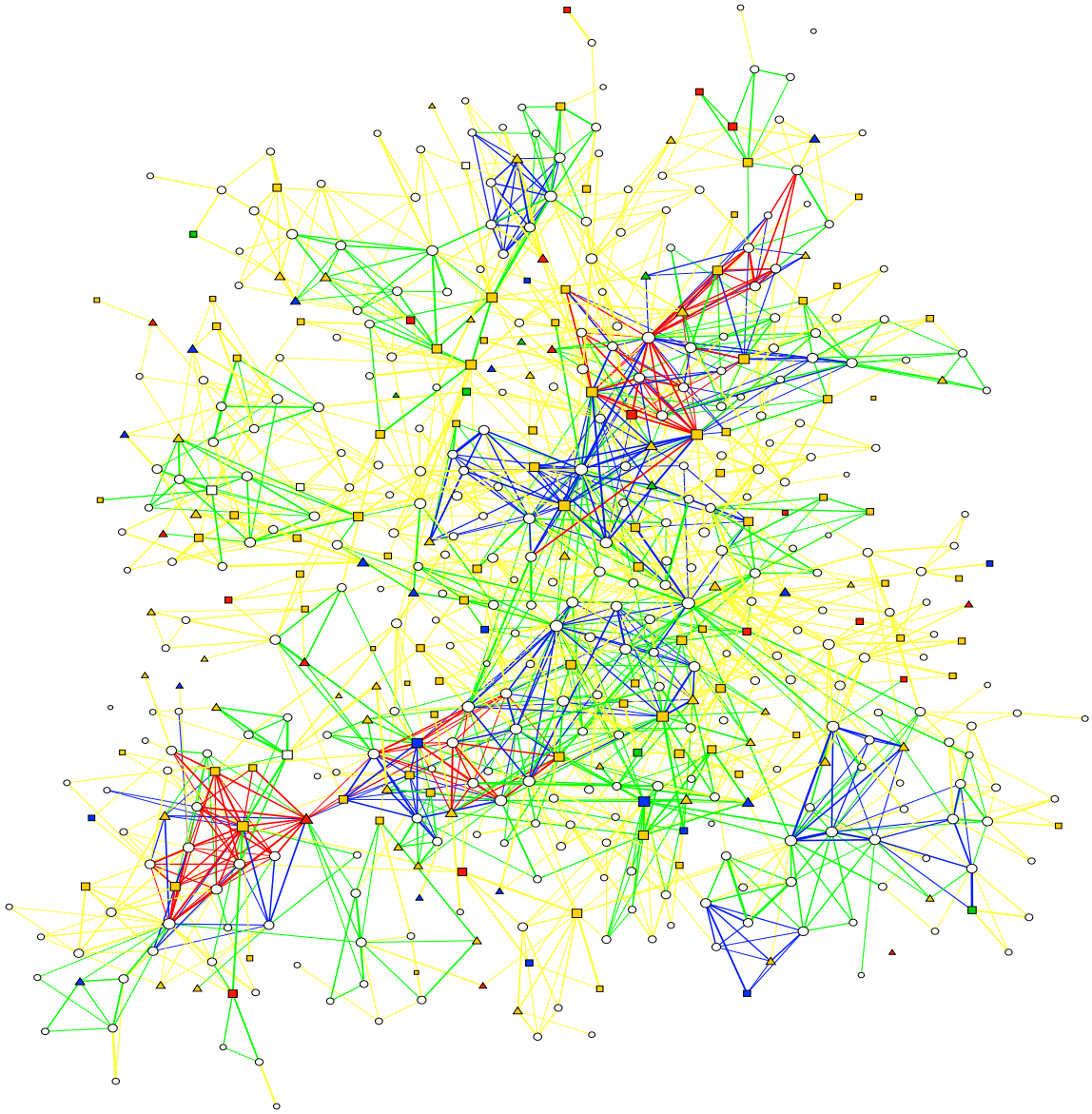


Figure (4.15): Plot of 4-7-clique communities in the largest 4-clique community. Red denotes the 7-clique communities. The 40 % weakest edges have been removed.

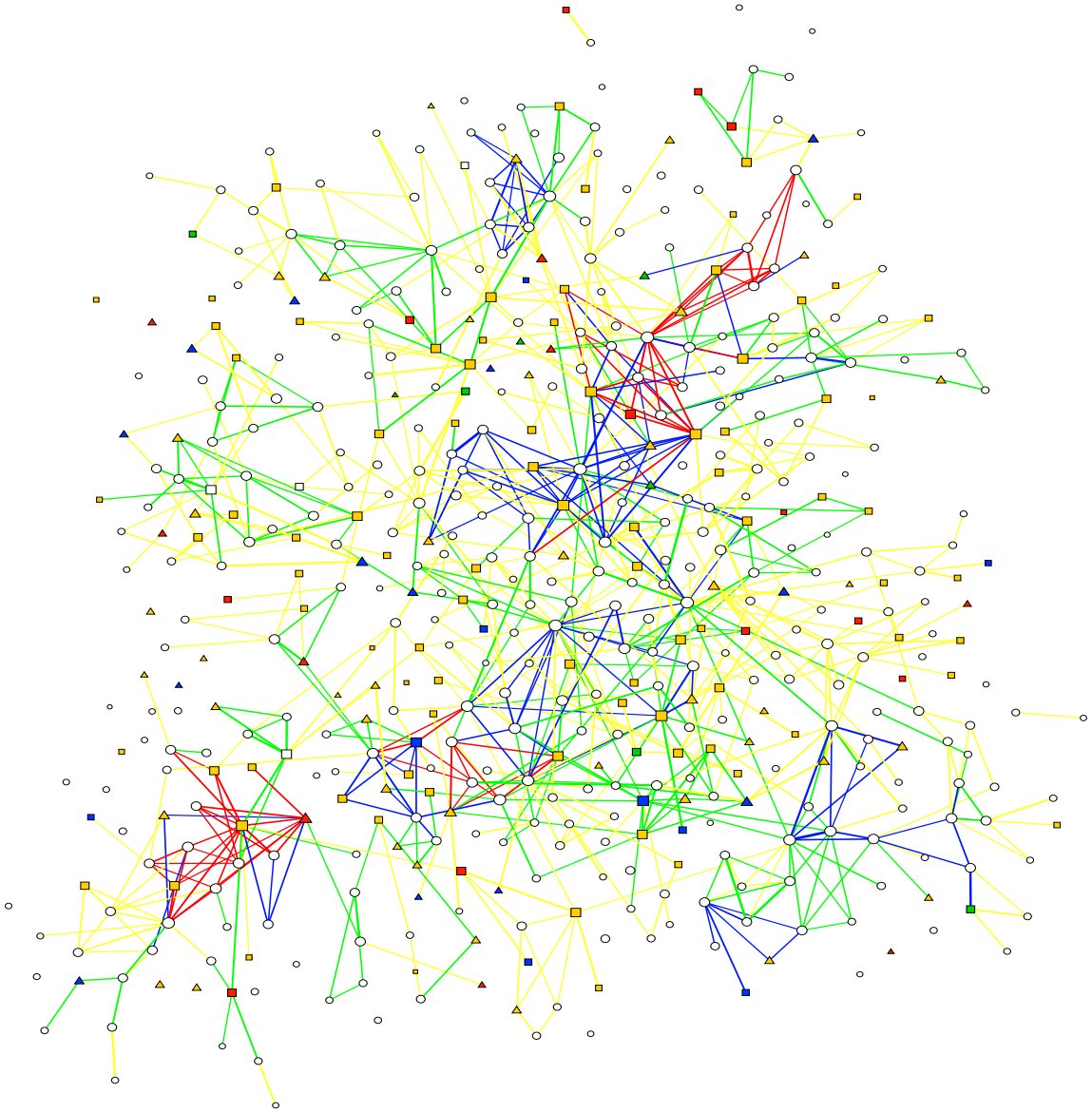
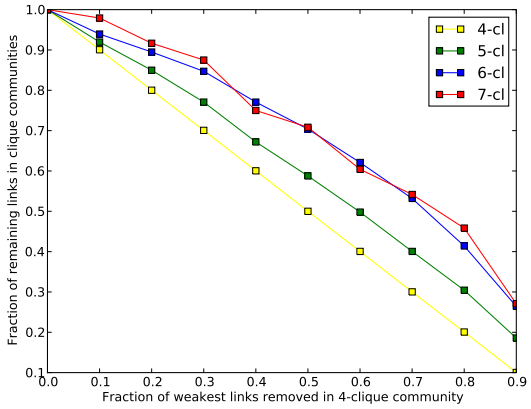


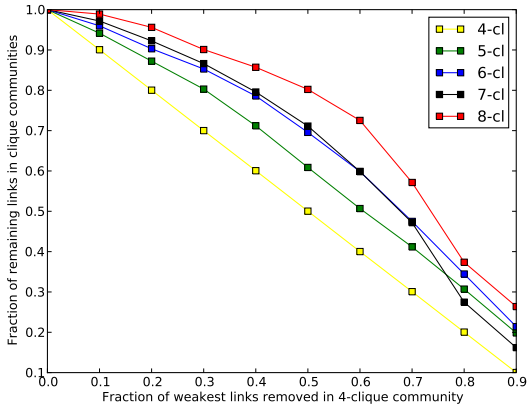
Figure (4.16): Plot of 4-7-clique communities in the largest 4-clique community. Red denotes the 7-clique communities. The 70 % weakest edges have been removed.

4.3.2 Thresholding analysis

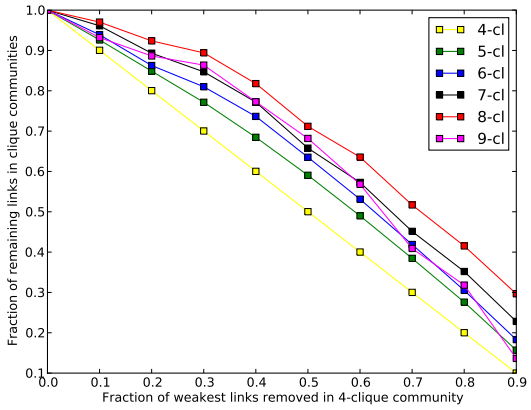
Here we present the numerical analysis of the edge weight thresholding for our four communities. The colors in the plots corresponds to the ones used on the edges in the visual plots.



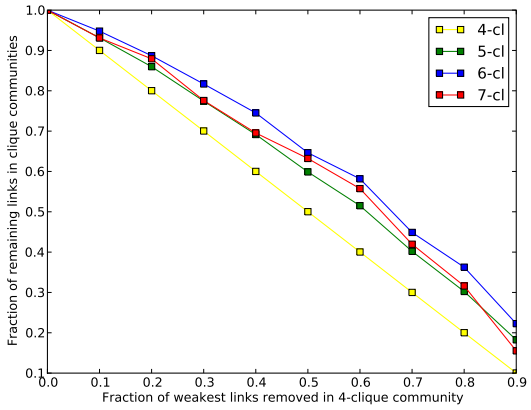
(a) 4th largest 4-clique community



(b) 3d largest 4-clique community



(c) 2nd largest 4-clique community



(d) Largest 4-clique community

Figure (4.17): Remaining fraction of edges in the different clique communities as the weakest edges is removed. The 4-clique line is linear since all the higher cliques are just even denser 4-cliques. We see that in general the higher k-value cliques resist the thresholding better. Interesting is it though, how the second highest (b) or highest (c) and (d) k-cliques seem to fall through for high thresholds in the three largest community plots. Unfortunately we have not done enough analysis to say if this is a very common trend for the highest k-clique communities.

Chapter 5

Numerical community analysis

Visualization does little to help us quantify the behaviour of communities. Therefore we have studied certain properties of the communities numerically. We have omitted the k-clique communities with a k higher than 8 from the plots as there are less than 30 of them all together. These communities can be said to be some extreme social phenomena. The highest k found in the network is 12 with 14 nodes in the community. Even for the lower k-values, we see in the plots that we would have liked to have more statistics for larger communities. We have divided the analysis in two parts, first we look at the structure and topology and then the demographics.

5.1 Structure and topology analysis

The following figures show the analysis of some of the properties discussed in chapter 2. These properties are not dependent on the demographics and are based on just the actual phone communication between the users. To give a better idea of the diversity of the communities we present both the scatter-plot and the binned average plot of the different properties studied. Note that we do not use the rather unpleasant color yellow in the following plots and have instead shifted the color scheme one place.

In figure 5.1 we first show the community size distribution. We see that there are "giant" communities forming for every k, containing a number of nodes much higher than the average. This is partly a percolation phenomenon, and if we could have used non-integer k-values we could have found the exact k for the percolation threshold between many small communities and one large one. This is also called a *continuous phase transition*, analogous to many phenomena in physics. Mentioned earlier was the giant community at 1.5 million people for $k = 3$.

The edge density naturally drops with community size, as shown in figure 5.2, and the average edge weight gets smaller when the community grows. The bigger the community gets, the looser connected people are. This is a bit interesting when compared to a study of dynamical communities. We have used the aggregated network over 18 weeks, but this study show that larger communities are more stable if there is a continuous exchange of

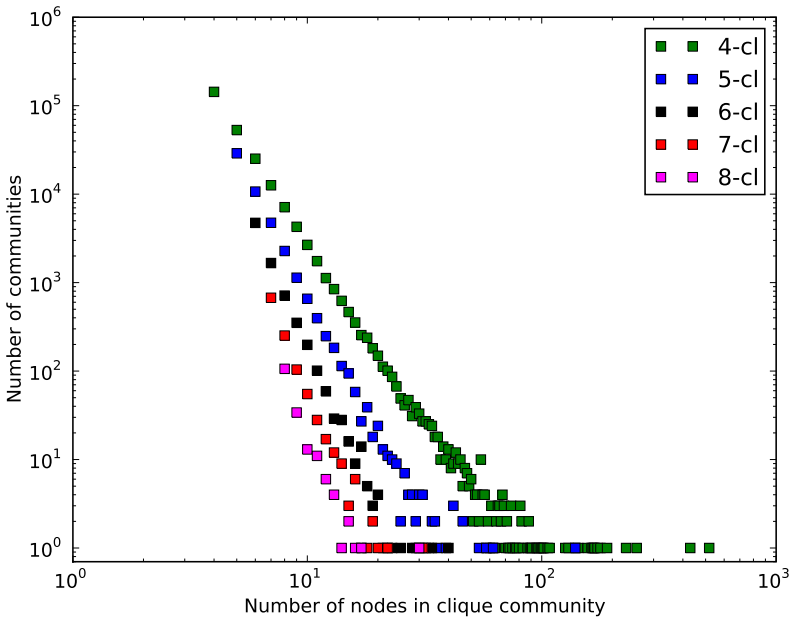


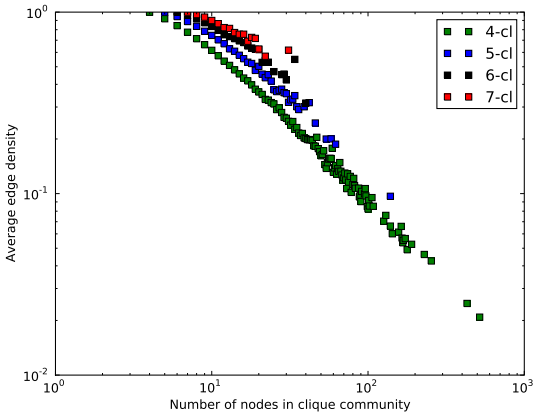
Figure (5.1): Community size distribution. (The straight line at the bottom is at one on the y-axis.) The distributions decrease sharply like a power law and the different k-cliques all have several "giant" communities with a number of nodes far above the average. The giants start to form at a fixed community size resembling a cutoff value.

members of the community. Small communities, on the other hand, are more stable if the members are the same in the whole period of study [24]. It might be that we have captured this observation in the plot of the edge weights, i.e., people have come and gone from the large communities without making a lot of calls. Even if these people have not called much, they have contributed with nodes and edges to our network.

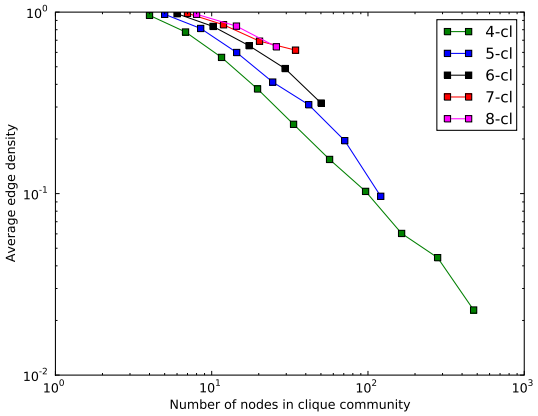
In figure 5.3 we see the average intensity and coherence. The intensity clearly drops for larger communities, but seems to flatten at some value. If we compare with figure 5.1, we see that the flattening appears to correspond to the community sizes where the mentioned giant communities start to appear. This seems to have some logic in the sense that those big communities are more random people not knowing all the others. Therefore the intensity of large communities is not as high as in smaller where people actually know each other.

The coherence has a very peculiar trend. We do have too little statistics in the tails of the plots to be certain, but we dare to propose that in both small and large communities people call more in an evenly fashion. From figure 5.2d we deduce that in small communities everyone call everyone more while in large communities everyone call everyone less.

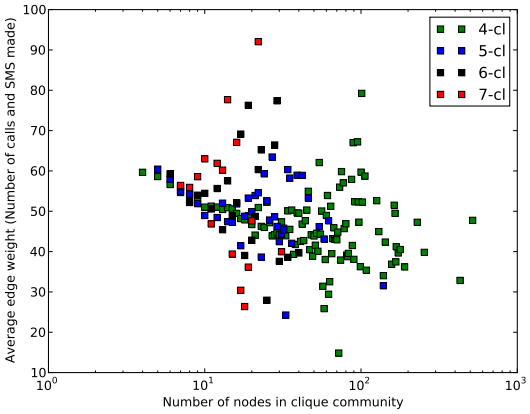
In figure 5.4 we see the average shortest path length and diameter. When we see



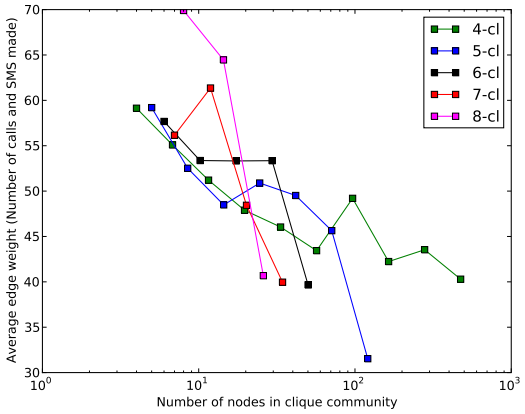
(a) Edge density, scatter-plot



(b) Edge density, binned plot



(c) Edge weight, scatter-plot

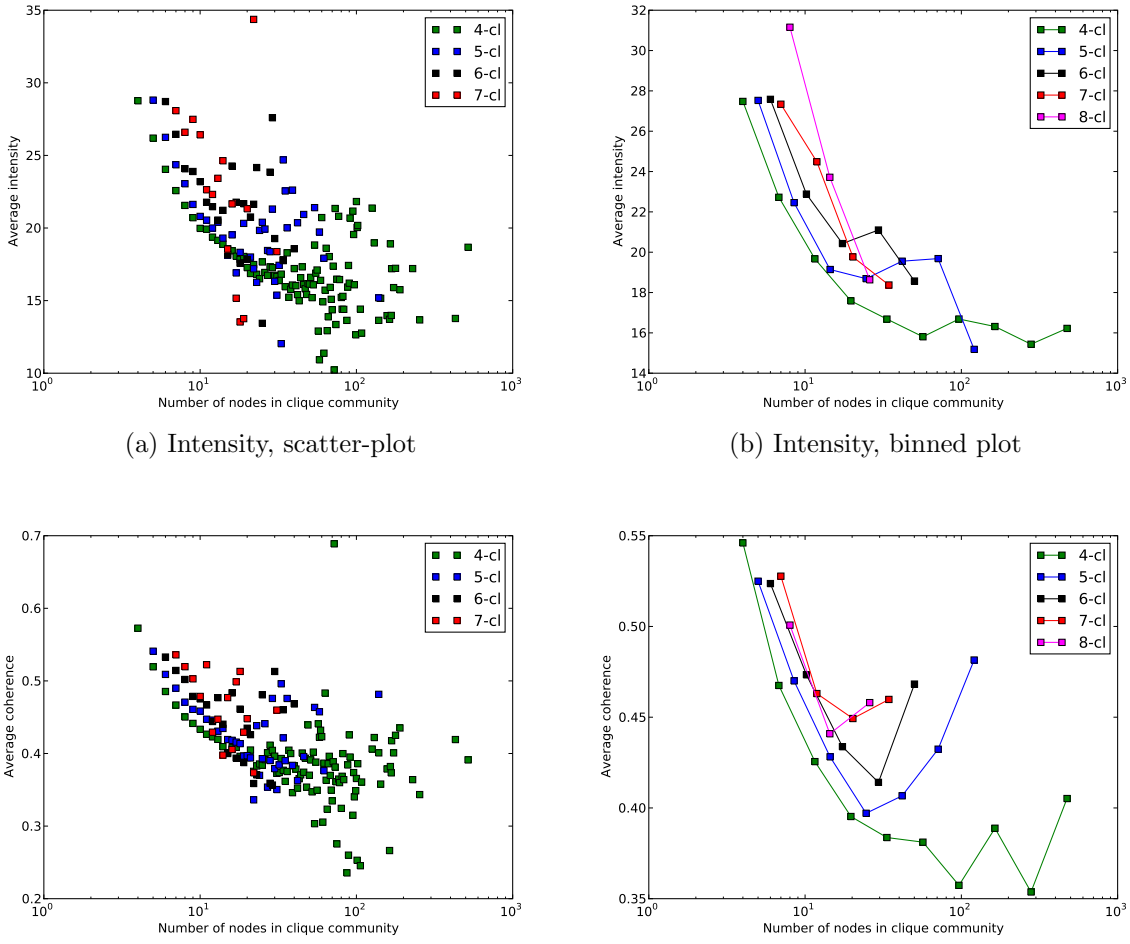


(d) Edge weight, binned plot

Figure (5.2): Average edge density and edge weight as a function of community size. (a) The edge density scatter-plot has a very unified shape, and a clear resemblance to a power law. The analogy here is that with our definition one does not have to know all the persons in a community personally to be a part of it, and therefore the edge density will naturally go down as the community grows. (d) For all k it appears clear that the average weight is stronger for smaller communities. One could think that in a large community one has to divide one's time between more people, so one makes less calls to each other person.

these plots, it is time to question whether a group of people with an average shortest path length over two (or diameter over four, for example) should be considered as a community at all. Going more than two steps in a community for finding someone with something in common seems a bit strange. Exceptions could be made for large companies, universities, big cities and so on, of course. However, there are only a hundred communities or so in

this range in the network, and they are almost exclusively 4-clique communities. This also raises the question if 4-cliques are the right basis to start from, or if some other measure to deal with the large diameter communities should be taken.



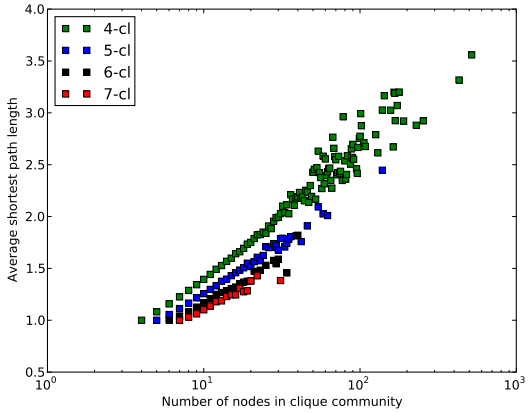
(a) Intensity, scatter-plot

(b) Intensity, binned plot

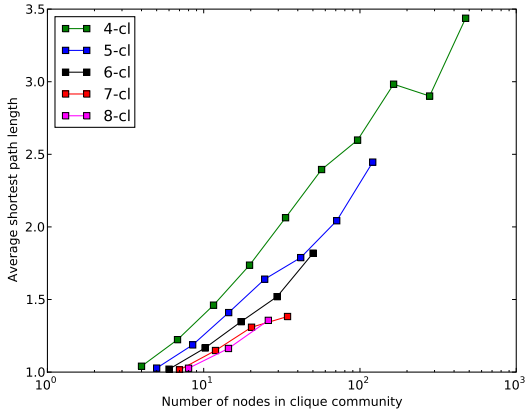
(c) Coherence, scatter-plot

(d) Coherence, binned plot

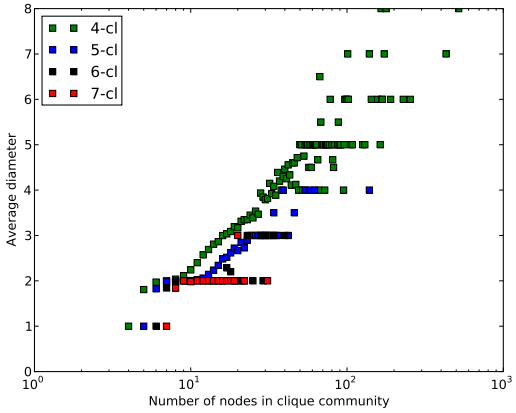
Figure (5.3): Average intensity and coherence as a function of community size. (b) The intensity is clearly decreasing for larger communities, but seems to flatten at the cutoff value mentioned in the text of figure 5.1. (d) The coherence seems to be highest for both small and large communities.



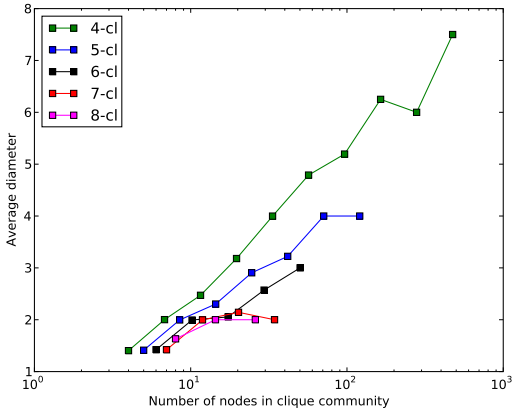
(a) Shortest path length, scatter plot



(b) Shortest path length, binned plot



(c) Diameter, scatter plot



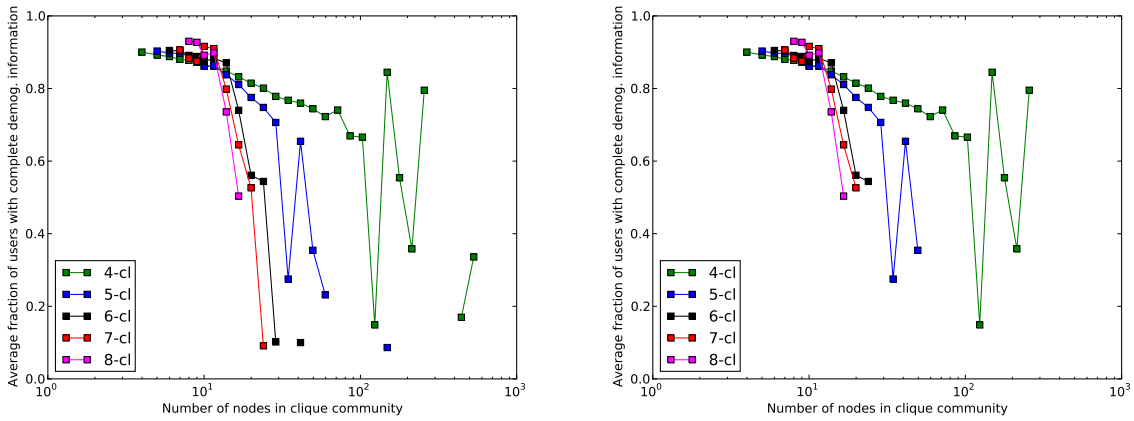
(d) Diameter, binned plot

Figure (5.4): Average shortest path length and diameter as a function of community size. (a), (c) We see that the shortest path length and diameter grows logarithmically quite uniformly. (b), (d) The highest k-values, 6 and 7, on the other hand, flatten for larger communities. This is quite naturally, since the increased number of edges required to form a clique shorten the paths and diameter of the community.

5.2 Demographic analysis

Here we take a closer look at the statistics of the demographic data available. As we saw in the visualization part of this thesis, the demographics have to be analyzed numerically. To ready the data for analysis we first looked at how many communities which had two or more users with complete demographic information. We found that only 1-2 % of the communities, at all the k-values we have included in our analysis, had less than two users with complete information. We decided that our qualitative understanding of the results

would not suffer from dropping these communities from the analysis all together. Next, we show in figure 5.5a the distribution of the average fraction of users with complete information versus community size in the remaining communities. From this result we decided to cut the largest communities at each k-value from our analysis. When the fraction of users with known information is below 30 %, combined with the very low number of such large communities, we simply can not know if any analysis will give us any trustworthy information at all. The updated data we are using for the further analysis is then shown in figure 5.5b.



(a) Initial data

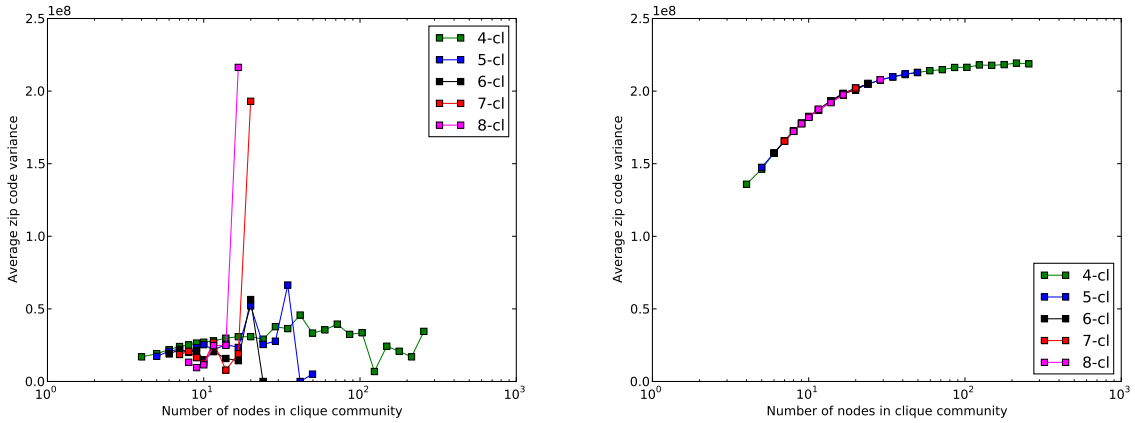
(b) Processed data used in further analysis

Figure (5.5): Average fraction of users with complete demographic information as a function of community size. (b) This is the data material we use in the further analysis.

For the analysis of age and zip code we used the average variance. Taking the variance of zip codes is clearly not an optimal solution, as the numbers scale with physical distance in a non-trivial way. From the data we have on the zip codes, however, we have verified that there is a system where ranges of numbers are reserved for geographical areas. This let us do a pure qualitative analysis compared to our network where the demographic properties of the users have been shuffled. We took an average of the variances over a thousand shuffles of the ages and zip codes and they are presented with the analysis of the original network. This kind of shuffling is common in network analysis, as one want to make sure that one is not just analysing something which is totally random. In figure 5.6 and 5.7 we see that our network is much more conform in the communities with respect to zip code and age than the shuffled reference.

We have done two separate analysis of the homogeneity of the communities with respect to sex. We have looked at the variance of the fraction of girls within each k-value, compared to one shuffling of the the sex of the users. And we have plotted the average largest fraction of one sex at each community size. As we see in figure 3.3, the number of men and women with known sex is not equal. Therefore we have to keep in mind that there is a possibility that our results are slightly biased because of it being four hundred

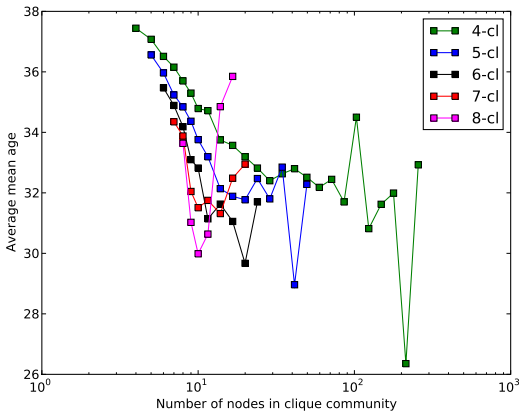
thousand more men than women in our network. In an ideal large social network there are probably equal amounts of both sexes. In figure 5.8 we see that the network is also more conform in the communities with respect to sex than the shuffled reference.



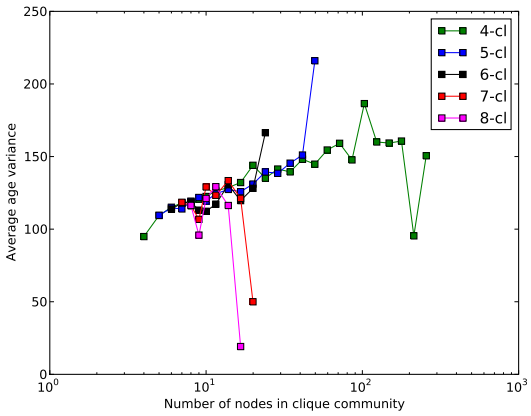
(a) Zip code variance, original

(b) Zip code variance, shuffled

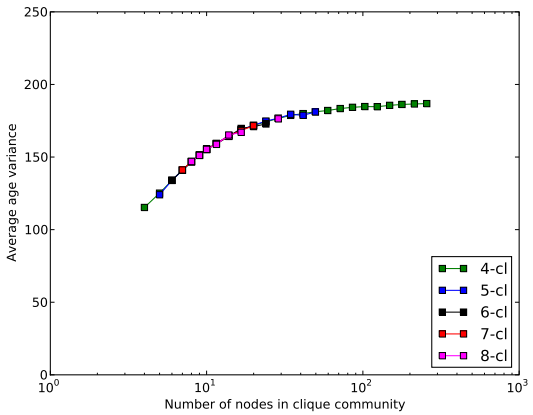
Figure (5.6): Average zip code variance as a function of community size. It is clear that people have more conform zip codes in real communities compared to the shuffled reference. Qualitatively, we do mean that it proves that people tend to know more people in the neighbourhood or at least in the same city, than elsewhere in the country. Why the largest 7 and 8-clique communities hold together over such long distances has to be attributed to one or more random factors.



(a) Mean age

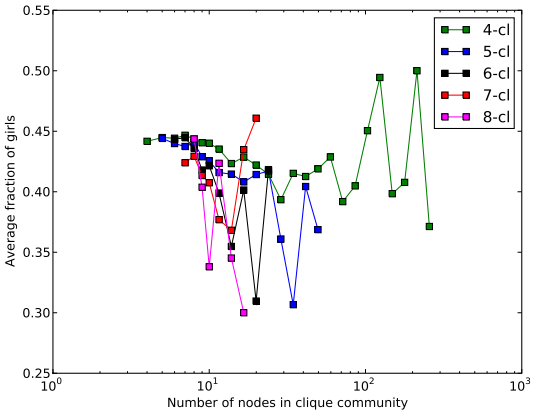


(b) Age variance, original

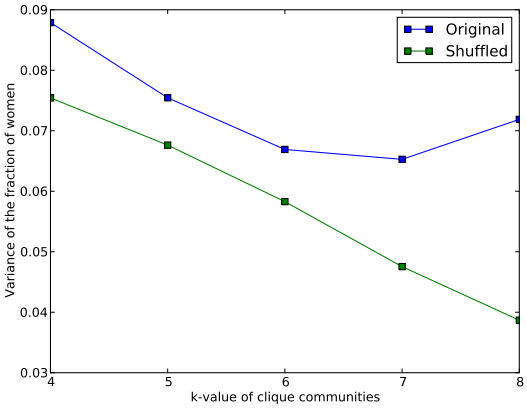


(c) Age variance, shuffled

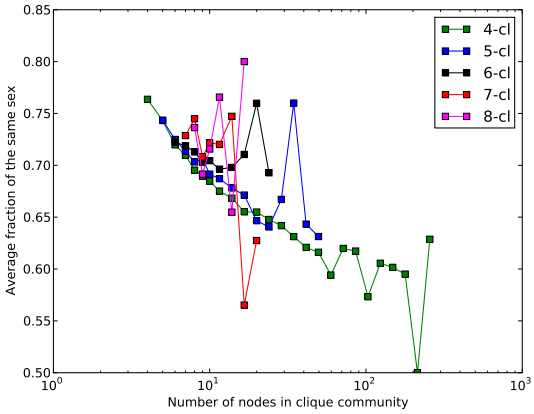
Figure (5.7): Average age statistics as a function of community size. (a) The average mean age is clearly higher for smaller communities before it drops and flattens. For the 7 and 8-clique communities the mean age seems to rise again for large communities, though. (b) The average age variance grows for larger communities, which is natural, with some random extremes for the largest communities. We see that the variance for the original network is clearly lower than for the shuffled in (c). This confirms the hypothesis that communities are more conform with respect to age than a random reference.



(a) Average fraction of women



(b) Variance of the fraction of women



(c) Average of the largest fraction of one sex

Figure (5.8): Statistics concerning the sex homogeneity of the communities. (a) The average fraction of women in each community reflect the fact that there are less women than men in our data, as seen in figure 3.3. (b) When taking the variance of the fraction of girls in each community, we see that the numbers are bigger for the original network than for the shuffled. That means that there are more communities with a fraction further from the mean in the original network, meaning again that there are more communities with a higher number of women than men and vice versa. (c) Taking the average of the largest fraction of one sex, we see that small communities tend to be more homogeneous with respect to sex than large ones. One could wonder if adding more women to the network would change the shape of these curves or not.

Chapter 6

Conclusions and further work

In this thesis we used the sequential clique percolation algorithm to detect communities in a large mobile phone user network. We then analyzed the detected communities and their properties both visually and numerically. The combined results of our analysis have given us reasons to recommend the use of the clique percolation method for detecting communities in social networks. The simple definition of a community leaves no doubt of what we have detected, and the detected communities show no signs of illogical behaviour in the results.

The high quality visual plots of the network let us clearly see two of the things we have been most excited about in social networks. First of all, we see nested communities, or a hierarchical community structure in the network. This indicates that we may be able to detect, e.g., different groups of friends within a workplace or university, or several families within a neighbourhood.

Secondly, our thresholding of edge weights shows that the strongest edges in the network are found within the densest sub-communities. This is our main result and both the visual plots and the numerical analysis give the same conclusion. We have also confirmed that weak edges keep different communities connected while the members of the communities have strong edges among themselves.

From the structure and topology analysis we have learned a lot about the average behaviour and calling patterns between the members of a community. Highlights include the fact that the average edge weight is less in large communities than in small ones. Also, the intensity drops for larger communities, and it has possibly a characteristic value for the giant communities which appear for all clique sizes. We also mention that the coherence seems to be equal in small and large communities.

From the demographic analysis we can conclude that all the commonsensical assumptions about the composition of communities are correct. We see that communities are more conform with respect to zip code, age and sex compared to the same network where the demographic properties have been shuffled.

6.1 Next steps

This thesis has only taken the first steps in the study of communities in social networks and there are several more topics to investigate. We have used the sum of the number of calls and SMS as weights in our analysis, but both duration of calls and the number of calls and SMS separately could also be checked. One could see some differences in behaviour and community composition with the different weights. The properties of the communities we have plotted as a function of the community sizes could also be plotted against each other to investigate any correlation between them.

As we saw earlier, our analysis suffered from the lack of complete demographic information of the users. In addition to what we know we would have liked to also know, e.g., the occupation of each user. It might also be a wise choice to do the same analysis for only the users with complete information in our network. If so, one have to lessen the already limited sample of mobile phone users we have. With this in mind, one is always on the lookout for new and more reliable data to do research on. If we could do the same analysis on other kinds of social networks we could be able to see some real patterns in social communities. E-mail, instant messaging or some of the increasingly popular online communities could all be the next source of a good data set. The process of getting access to such a data set as we have had is lengthy and laborious, however, and it is difficult to find data sets with even better demographic information than the one we have already investigated.

Bibliography

[1] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 915–924, New York, NY, USA, 2008. ACM.

[2] Wayne W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(5439):452–473, 1977.

[3] Aurelien Gautreau, Alain Barrat, and Marc Barthelemy. Microdynamics in stationary complex networks. *PROC.NATL.ACAD.SCI*, 106:8847, 2009.

[4] Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 435–443, New York, NY, USA, 2008. ACM.

[5] Albert-Laszlo; Barabasi and Reka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[6] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393, 1998.

[7] J. P. Onnela, J. Saramaki, J. Kertesz, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71:065103, 2005.

[8] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[9] Renaud Lambiotte, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317 – 5325, 2008.

[10] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, Apr 2001.

[11] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378, 2000.

[12] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626–4628, Nov 2000.

[13] Reuven Cohen, Shlomo Havlin, and Daniel ben Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91:247901, 2003.

[14] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002.

[15] A. Arenas. Community analysis in social networks. *The European Physical Journal B - Condensed Matter*, 38:373–380(8), March 2004.

[16] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35:66–71, 2002.

[17] Pall Jonsson, Tamara Cavanna, Daniel Zicha, and Paul Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, 7(1):2, 2006.

[18] Jussi M. Kumpula. *Community structures in complex networks: detection and modelling.* PhD thesis, Technical University of Helsinki, Finland, 2008.

[19] Jussi M. Kumpula, Mikko Kivela, Kimmo Kaski, and Jari Saramaki. A sequential algorithm for fast clique percolation. *Physical Review Letters*, Jul 2008.

[20] Imre Derényi, Gergely Palla, and Tamás Vicsek. Clique percolation in random networks. *Physical Review Letters*, 94(16):160202, Apr 2005.

[21] Mark S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[22] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabasi. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[23] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9:179–+, June 2007.

[24] Gergely Palla, Albert-Laszlo Barabasi, and Tamas Vicsek. Quantifying social group evolution. *Nature*, 446:664, 2007.