Structural difference between public and private communication in an online community

Fariba Karimi
Physics Department, Lund University
Supervisor: Petter Holme
Umeå University

May 17, 2011

Abstract:

We investigate an online community where people meet and communicate in forms of discussions or sending messages to each other. We try to see the big picture of communications and social activities by network analysis. Specifically, in this social network, there are two modes of communications. Either a user can reply to others in a public forum in such a way that we see who comments on whom; or they can send e-mail-like direct messages. In this data we investigate network structures (such as degree-distributions and assortativity), temporal structures such as response-time, interevent-times and activity levels. Furthermore, we measure combined structures from the different communication channels relating to social-balance theory. Among other things, we find that in private communication, people keep feeling obliged to reply over a longer period of time, than in public discussions. We also observe a weak anti-correlation between activity levels in public and private communication respectively, suggesting that different personality types drive the large-scale structural evolution. We relate our findings to theories of social organization and human dynamics.

Sammanfattning (summary in Swedish)

Modern kommunikationsteknologi genererar dataset som kan hjälpa samhällsvetenskapen att förstå hur vi kommunicerar: när vi skickar meddelanden, till vilka och hur vi organiserar våra sociala nätverk på nätet. Storleken på kommunikationsdataset gör att man måste använda enkla mekanistiska modeller av samma typ som förekommer i statistisk fysik för att förstå ett dessa system. Detta arbete utgår från ett data-set om kommunikation på ett Internet community. På detta community kan människor mötas och kommunicera både genom diskussioner på ett forum och e-mail liknande meddelanden. Jag använder nätverksteorier för att beskriva den storskaliga strukturen av kommunikationen och försöker dra slutsatser om de sociala processerna som ger upphov till dessa strukturer. I detta community förekommer det två typer av kommunikation: användare kan dels diskutera i ett offentligt forum på ett sådant sätt att vi kan se vem som kommenterar vem, eller så kan de (som sagt) skicka privata meddelanden. Tack vare dessa skillnader i kommunikationskanaler kan vi dra slutsatser om vad som skiljer kommunikation i ett privat och offentligt sammanhang. I detta data mäter vi nätverksstrukturer (t.ex. sannolikhetsfördelningen av grad—antalet samtalspartners—assortativitet, d.v.s. tendenser att personer med många kontakter mest kommunicerar med andra med många kontakter, och de med få kontakter kommunicerar med de med få kontakter), temporala strukturer (såsom svarstider, tider mellan ivägskickade meddelande och aktivitet). Dessutom mäter vi kombinerade strukturer från olika kommunikationskanaler och relaterar dessa till teorin om social balans (en sociologisk teori som säger att konfigurationer av tre personer med två positiva vänskapsband och ett negativt band av två personer som ogillar varandra kommer vara instabila). Bland annat finner vi att i privat kommunikation, har personer en större benägenhet att svara snabbt, jämfört med offentliga diskussioner. Vi observerar också en svag negativ korrelation mellan aktivitet i offentlig och privat kommunikation och relaterar detta till att bland de aktiva användarna finns det olika personlighetstyper som driver utvecklingen av communityt.

Contents

Chapter 1

Introduction

In this chapter we discuss general concepts which are related to the aim of this project and social networks.

1.1 Complexity

"Ideas thus made up of several simple ones put together, I call complex; such as beauty, gratitude, a man, an army, the universe."

John Locke, English Philosopher (1632-1704)

If we look around ourselves, despite simple physics rules we have learned at school, nature does not look that simple. The reality is more complicated. Sometimes these complicated patterns still have relatively simple explanations, for example snowflakes, mountain ranges, ridges on sand dunes or even financial markets. The question is: if the laws of physics are simple, why are natural systems so complicated [9]? One methodology of understanding how a system works, is to separate parts of the system and study each piece one by one. Physics and generally natural sciences, try to simplify and separate the system to be able to explain how the system works. For example, in high energy physics, scientists are trying to find the most fundamental elements of the universe. But then what kind of understanding do we have? Can we take the knowledge from particle physics to understand why bacteria behave the way they do? Well, not without putting the elements back again and look at the whole system.

Many natural systems cannot be understood by their structures only. One has to keep functionality in mind. Concepts like growth, reproduction and adaption are hard to define without their functions in a system. A dead body is still complex but not functional [5]. Therefore, to be able to understand natural systems, we need to consider elements of a system, but in connection to other elements and their functionality.

One ingredient in *complex systems* in physics is *"emergence"*—that properties of a system consisting of many interacting units come from the multitude of

interactions rather than the properties of the individuals: "More is different" [2]. This is usually defining the difference between complex and *"complicated"*. Complexity arises from the consequences of collective behavior not the complicated properties of individuals [26].

Another concept related to complexity in physics is *chaos*. In chaotic systems, results depends on initial conditions. Errors and unpredictability grow exponentially as time passes by. In chaotic systems one cannot predict the future exactly, not because of noise, but the sensitivity of the mathematics behind the system. Sensitivity to initial conditions can thus produce another type of complex behavior system, meaning that systems can still obey simple laws, even if they are complex

To be able to study complex systems, it is important to consider the right level of description. Some large-scale properties do not depend on small-scale details. Thus by using the right level of description, one can understand the complex system. *"Don't model bulldozers with quarks" [9].*

In recent years many interdisciplinary fields have been emerged by studying complex systems. One example is in financial market. In financial systems people, traders, producers and stock markets are competing and cooperating in such a way to get more benefits compare to others. Such systems have been studied using tools in physics, in the new research field, econophysics [16].

1.2 Statistical physics

Statistical physics is a branch of physics that explains thermodynamic properties in mechanistic way. It provides a general framework for how macroscopic properties of the system emerge from microscopic variables. Since the framework is successful and general, it has been applied to other fields of science such as biology, medicine, computer science and even sociology. In this project we use methods from statistical physics to study an online social network.

1.3 Statistical physics and social networks

In a society, people interact and communicate with each other. The interaction could be in the form of friendships, sending emails, trading, sexual contacts and so on. This interaction is complicated to characterize and measure. It is also interesting to look at macroscopic phenomena which are results of the many interactions of individuals in a social system. Examples of the transition from behavior of individuals to large scale phenomena in social systems, could be the emergence of new language, culture, art, etc. Thanks to the fast growing online social networks, scientists get source material to understand such big scale phenomena quantitatively.

There are, however, difficulties in applying statistical physics to social systems. In mainstream physics elements of a system are well defined and their behavior are well known. Therefore, by understanding the microscopic details and

using probabilistic methods, one can, by statistical physics, derive the macroscopic properties of the system. However, in social systems, details of people's behavior and microscopic variables are not so well defined. Therefore, modeling of large scale social systems is not so easy. Macroscopic outcomes are also harder to understand and interpret.

What makes statistical physics a useful approach to study social systems is the fact that many macroscopic properties are independent of microscopic details. One example of macroscopic property are statistical physics is *universal* exponents. Imagine a container of gas. Close to phase transition interesting things are happening. For instance heat capacity, C, close to phase transition, shows power law behavior as temperature changes:

$$C \propto (t - t_c)^{\alpha} \tag{1.1}$$

Where t is the temperature and t_c is the critical temperature and α is a critical exponent. By changing the temperature in the system, t_c varies but α remains constant. Other physical quantities of the system are also following same pattern (scaling behavior), close to phase transition. Interestingly, other critical exponents are related to each other. This phenomena also happens in ferromagnetic phase transitions in magnets. This is an example of a universal quantity of a system remains constant, although microscopic details of the system are changed [14]. Another concept in statistical physics is phase transitions between disorder to order states. Imagine a two dimensional lattice with small magnets distributed evenly over the lattice. Each magnet tends to arrange it's poles depending on it's neighbors in such a way that the opposite poles are closer to each other than poles of the same sign. Eventually by the influence of neighbors, the whole system will have magnetic properties. In sociology point of view, individuals tend to share the same opinion, culture, language,... as time pass by. The evolution of social states during time can be characterized by statistical physics tools. In social science literatures, the tendency towards similar behavior between people in the society is called *social influence*. The counterpart of social influence in physics could be parallel alignment in ferromagnetic materials.

1.4 Empirical studies of social networks

There are many questions that are interesting to ask about human behavior such as how we schedule different tasks, make decisions, create and develop new ideas. These phenomena are a consequence of physical constraints, biological factors, social effects and so on. The most fundamental question in sociology is why there are differences in society. Some aspects of that question, but not all, could be addressed by statistical physics methods

An early study of social networks by physicists was about an email exchange network conducted by Ebel et al [7]. They studied email exchange network

of students in a university server[1]. They observed scale-free distribution in the email network and also discussed the spreading of computer viruses in the network. In another study by Newman et al. [22] the authors observed the similar structure in email network and discussed about how to protect a network against email viruses.

In 2004 Eckmann et al. [23], observed the emergence of self organized pattern in email traffic. They showed how the pattern of sending email and getting response looks like between two people in the network in terms of time. They also discussed email exchange between 3 people in the network (triangles). They showed the emergence of making dialog between people and clustering in the network. It worths mentioning that there is a difference between Ebel and Eckmann observations regarding to email exchange. One considered both emails from inside the network and outside, while the other study only considered emails inside the system. Regarding to the data set, the email exchange is only happening inside the system (Filmtipset.se) and no one except members are able to send a message in public forum or private messaging.

The idea of *burstiness* is developed by computer scientists Jon Kleinberg [13]. He noticed that the frequency of using words is related to the importance of the up-coming topics. While a new important topic pops up, the frequency of related words are suddenly increasing. He also developed an algorithm to detect for burstiness in the network. Later on Barabasi explained the bursts effect and probability distribution of the email response time by using decision-based queuing process [3]. In another study, Holme et al. [11] analyzed data from a Swedish online dating community. They considered time evolution in the study of the network. Time-resolved network studies, or longitudinal studies, as they are called in sociology, are significant feature of study of networks in future[21].

We end up this chapter with an interesting characteristics of human's network, power law distribution.

1.5 Power-law distributions

One popular topic in the studies of complex networks is the emergence of power law behavior. Probably you have heard about stories like majority of the wealth in the world is in the hand of minority of people. If you imagine the spread of the wealth in the big network of people, one could see that it is not randomly distributed but few people (nodes) in the network have lots of money and many others do not. Surprisingly many real-world networks are known to show such a power-law behavior or at least they are close to it. Even if you look at the picture on the home page of online social network facebook.com, you will notice the image, represents few people with many links and the rest with less links.

[1] They only considered emails which are exchanged by internal accounts in the network. They ignored emails that users received or sent from/to outside of the server. In our case there is no possible to get an email or message from outside the community.

In 1896 economist Vilfredo Pareto first mentioned the logarithmic pattern of distribution of income and wealth [6]. In 1932, the American linguist George Kingsley Zipf, studied the frequency of English words [31]. He found out that the probability of using a word is exponentially increasing by its occurrence. In the other word if the rank of a word is k, then the frequency is proportional to $1/k^{\alpha}$ where α is close to one. This is known as *Zipf's law* or Zipf's distribution.

The current interest in power-law, or scale-free, networks started in 1999 by Albert-László Barabási et al. They Showed the emergence of power law scaling in the networks [4]. They showed in different networks such as actor collaboration, world wide web and power grids similar power law pattern in the network development. They also introduced a mechanism which cause the emergence of scaling in random network. The mechanism suggested that the network is continuously growing by adding new vertices and also new vertices are likely to attach to highly connected vertices later on known as *preferential attachment*. The power-law fit for the degree distribution, is also observed in many other networks such as the Hollywood graph of actors if they appeared in the same movie, power-grid network or phone-call network. If we assume power-law behavior in many real-world networks as an inherent property of the network, the question is why and in what extend it can be correct statement? To what extent could be the power-law distribution of a network a signature of self organizations? In chapter 4, we show the degree distribution in our data set and compare it with power-law properties.

Chapter 2

Preliminaries

In this section we introduce concepts and measurement methods which we will use later on in the result chapter, chapter 4.

2.1 Basic concepts in Graph theory

We will let *Graphs* denote the mathematical objects that represent *network* in real systems. Some example of special graphs are shown in Figure 2.1.

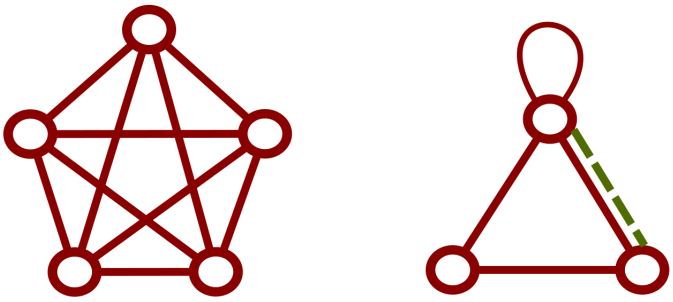


Figure 2.1: Sample of two graphs: a) A complete graph. Each pairs of vertices are connected by an edge together. b) A multigraph with multiple edge and a self loop.

Understanding the mathematical tools of graph theory, will help us to study networks in general. We will try to explain few concepts which we use later on in this project. Consider a simple directed graph like Figure 2.2.

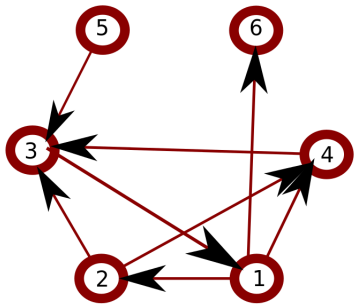


Figure 2.2: A simple directed graph

Vertices, in Figure 2.2 are red circles which are connected by red wires together. They are known as *nodes* in computer science, *sites* in physics and *actors* in sociology. In this thesis nodes are people who are members of the movie community and present in our data-set. Therefore whenever we refer to *member* in the network, we mean a node in the network. *Edges* in the graph are red wires who connect vertices together. In other literature they are also known as *links* or *bonds* or *ties* between vertices. In this project, edge represents a connection between two members in the data-set, defined by the sending or receiving a private Message or posting comment to each other in public Forum. We should mention that in the data set each edge consists of *attribute* or extra information which shows the time of messaging. Thus each edge consist of a list of time while node i sends a message to node j for example. A graph, G, with V vertices and E edges represents as $G = (V, E)$, in this example $G = (6, 8)$.

Graphs can be *directed* or *undirected*. In the Figure 2.2 the graph is directed since the edges show not only connection between nodes but the direction by small object in one end. For instance in directed graph, an edge (i, j) which shows the connection between two person i and j by sending message from person i to person j. Thus we get extra information about who is sender and who is receiver in the network. On the other hand sometimes we don't need to care about the direction, then we simply consider undirected graph. Another useful mathematical tool in graph theory is *adjacency matrix*. If the edge (i, j)

shows the joint of vertices i and j, it is adjacent or connected. The connection between edges are shown as A_{ij}. The value 1 shows a connection, while if there is no connection between two vertices i and j, then $A_{ij} = 0$.

Here we will explain concepts which are more advanced. *Clique* is a subgraph (a part of a graph), which every pairs of it are connected. Going back to the Figure 2.2, there is a clique of four between nodes $1, 2, 3, 4$. A path in the graph which ends to the same node as it starts is a *circuit*. A circuit with three edges, or a clique of three vertices, is *triangle*. We use triangles to check structural balance in the network[1]. Number of links which are attached to each node is a *degree* of the node. For example in Figure 2.2 node 1 without considering the direction has a degree of four. In directed graph, links which are pointing out from a node are counted as *out-degree*. In Figure 2.2 node 2 has out-degree of the order of two. If links are pointing towards a node they are counted as *in-degree*. In Figure 2.2, node 2 has one in-degree links.

In the next following sections we will discuss about metrics and tools which are closely related to the measurements in chapter 4.

2.2 Degree distribution

One important characteristic of the network is the degree distribution. Since an edge, in general, is used to indicate the possibility of interaction (and certainly in this Thesis this is the case), it is informative to know with how many others a vertex interacts. By definition, the degree distribution tells us how frequent the degree is in a network. It is known as one of the most fundamental properties of a network [21].

The degree distribution can tell us something about the evolution of the network. It can also affect dynamic systems on the network. For example, in internet traffic, the speed of information transfer, is affected by the form of degree distribution. In many real-world networks, the pattern of degree distribution is such that there are many nodes with low degree and few nodes with high degree. The nodes which have high number of edges connected to them are also known as *hubs* of the graph. For example in the network of Internet servers, most of servers have only few numbers of connections to other servers. But there is most highly connected server with 2407 connections, which means it is connected to 12% of all other servers in the network[21]. One other example is spreading HIV in the network of people. The degree distribution of the network of people in terms of their sexual contacts is not randomly distributed. Few people have many sexual connections such as prostitutes or sex-buyers and many other people have few sexual contacts. This distribution of sexual contacts, affects dynamic of disease spreading. There are many other examples of human-related networks which also have similar degree distribution characteristics. Degree distributions has been a hot topic since Barabasi's and Albert's paper in 1999 [4]. They showed that in many networks such as WWW or scien-

[1] See section 2.6

tific citation network, the vertex connectivity [2], is following a scale-free pattern. They showed that large scale networks tend to scale-free pattern in their degree distribution. They showed two factors which cause this self-organization: 1. The continuous expansion of the network due to adding new nodes or vetices 2. Vertices attach to other vertices which have large number of connectivity[3]. One example of similar scale-free distribution in physics is fractals. This self-similar properties of the network made scientists interested to study more deeply about different networks and their properties. Since Barabasi and Albert's paper[4], networks with power-law degree distributions have been known as scale-free networks.

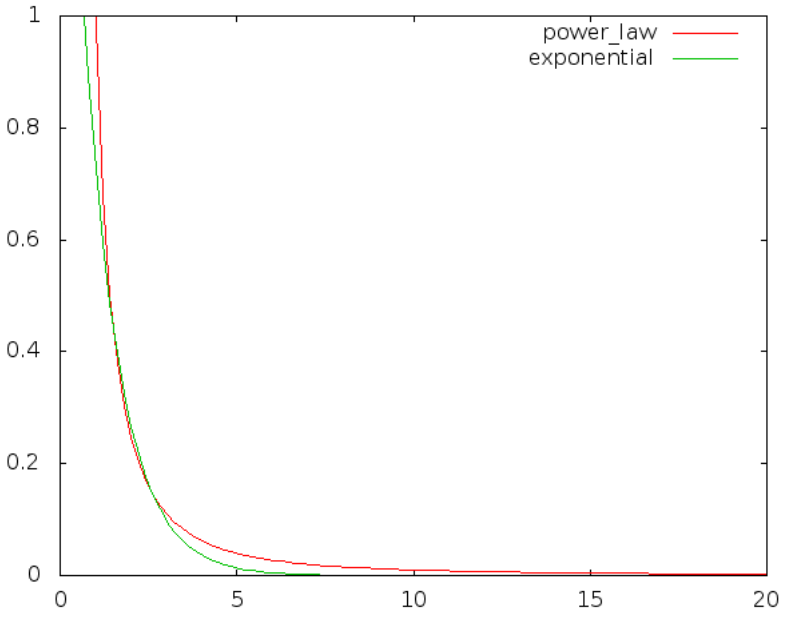
Knowing the degree distribution of the network is important. Imagine the example of disease spreading in the network. Vaccination of a whole society is a major concern of governments. It is costly and not possible to immunize the whole population due to limited resources. Understanding the network of how the disease spreads, could help to find better strategies to prevent epidemics. For example one valuable information which came from studying the network of HIV spreading is that people you have met recently are more likely to be socially active and thus central in the contact pattern, and thus important to vaccinate [15]. Another example is the network of power grids. In case of damaging in the network or sudden black-out, one could detect problem faster by knowing how the distribution of grids look like.

In the next section we will explain different methods to measure power law behaviour in the network.

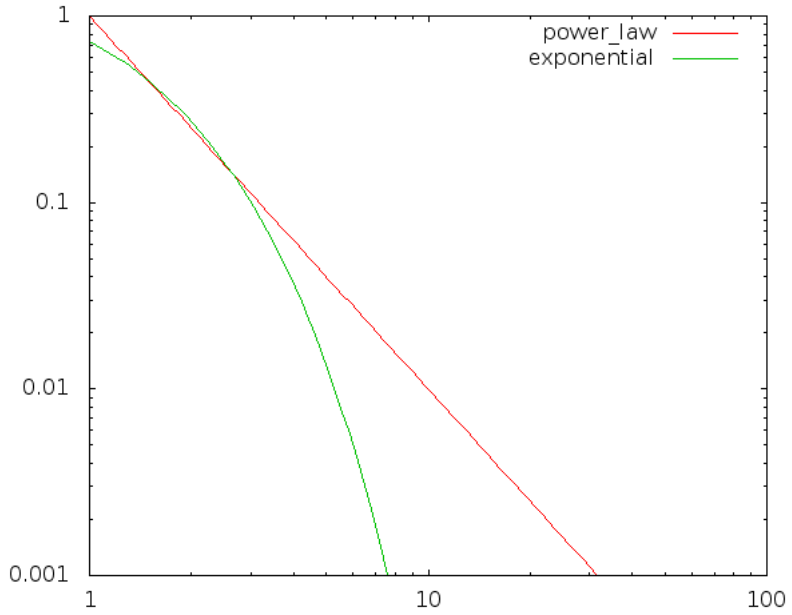
2.3 Measuring power law distribution

Detecting power-law distribution from the normal plot is not trivial. There are other similar functions which look like power law distribution in normal plot. Figure 2.3, shows two function with similar curve in normal plot: Exponential $p(x) \propto \exp^{-x}$ and Power law $p(x) \propto x^{-2}$ while in log-log plot they are distinguishable.

[2] Which later on known as degree distribution
[3] This is known as *preferential attachment*



(a) Power law and Exponential functions in normal plot. They look very similar and hard to distinguish from each other.



(b) Same two functions in log-log plot. Exponential function decrease faster than power law.

Figure 2.3: Comparing two functions in normal plot and log-log plot.

Therefore in the next section we will explain three methods to be able to detect power-law behavior more accurate.

2.3.1 Log-log plot

Plotting in log-log axis, as illustrated in Figure 2.3, is fast but not very accurate. At least it is good to check quickly if we are close to power-law distribution or not. Consider again a power-law function:

$$p(x) = Cx^{-\alpha} \tag{2.1}$$

By taking a logarithm from both side we get:

$$\ln(p(x)) = -C\alpha \ln(x) \tag{2.2}$$

In logarithmic plot the power law function is identified by a straight line with a slop proportional to exponent α. As an example, we take our data set of Forum pages from the movie community and plot degree distribution in log-log plot. See Figure 2.4.

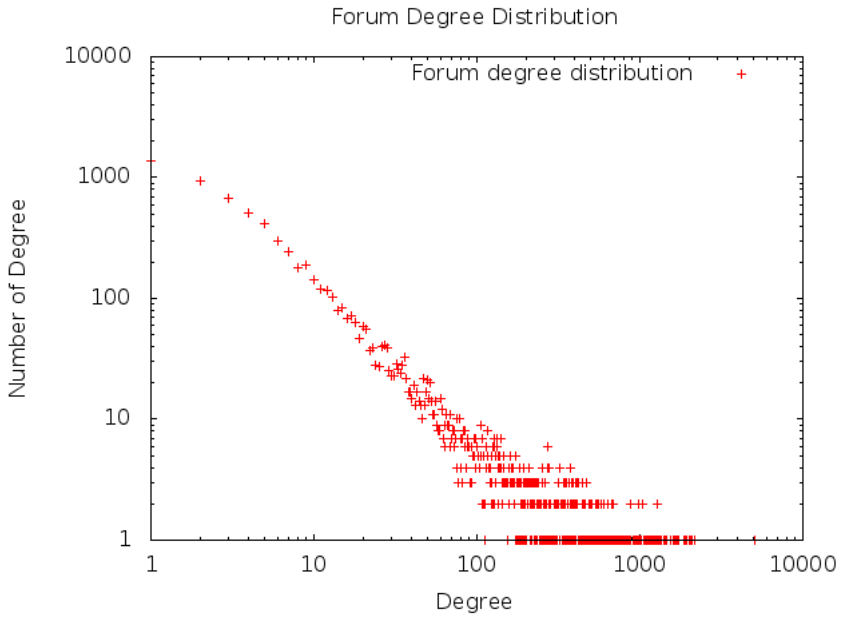


Figure 2.4: Forum posts degree distribution plotted in log-log axes.

One problem with log-log plot is that since there are less data in tails of the graph, the tail looks noisy. The noisy or fat tail makes it difficult to accurately

predict the distribution function of the system. Additionally, taking away valuable data from tail is not a wise idea. The tail-events are the most dramatic, and often most important.

2.3.2 Logarithmic binning

To get rid of the noise in the tail, logarithmic binning (or logbinning) is helpful. In this method, instead of making bins of equal sizes and count data in each bin, the bins themselves are getting larger and larger when the number of data decreasing (i.e. for large values of the quantity). So that there will be more data counted in the tail and noise will be reduced. After counting data in each bin, they are divided by the width of the bin so that the plot is independent of the binning method. Figure 2.5 shows the same data which are plotted by logbin method.

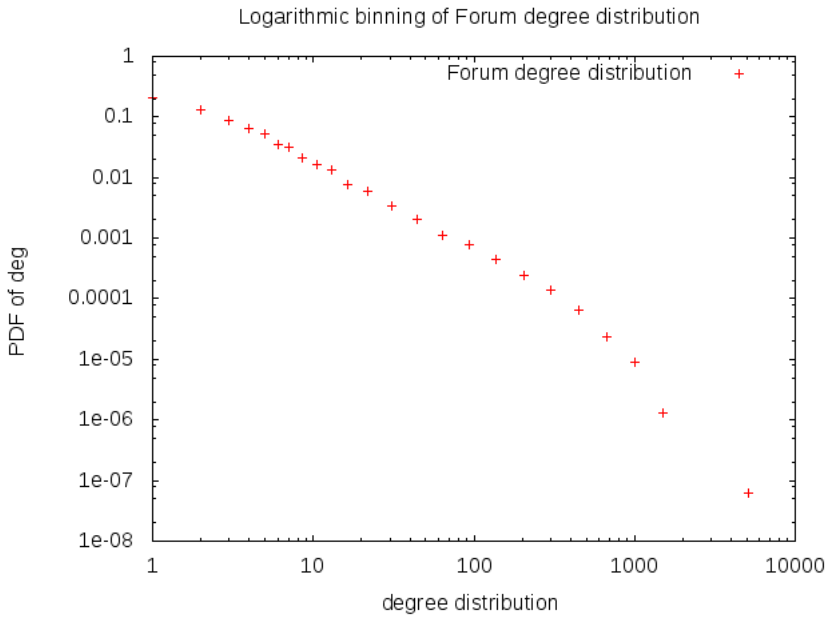


Figure 2.5: Forum posts degree distribution plotted with logbin method.

logarithmic binning reduces the noise in tails significantly, but not completely. We will explain third method which is more recommended than logarithmic binning [20].

2.3.3 Cumulative Distribution Function (CDF)

A better method for measuring power law is using Cumulative Distribution Function (CDF). Here we are not looking at the histogram of the data itself,

but the probability of each point in the graph is equal or larger than that point. Considering the same power law function, equation 2.1, in CDF method we sum up all probabilities equal and larger than x:

$$P(x) = \int_{x}^{\infty} p(x')dx' \tag{2.3}$$

Using integral in this case make sense because the distribution is varying very slowly. By taking summation of Equation 2.1:

$$P(x) = C \int_{x}^{\infty} (x')^{-\alpha} dx' = \frac{C}{\alpha - 1} x^{-(\alpha-1)} \tag{2.4}$$

Therefore P(x), cumulative distribution function, is a power law function with exponent $\alpha - 1$. In Figure 2.6, we plotted the same results for degree distribution of Forum posts but with CDF method. Here we completely get rid of noise and the same time keep all information.

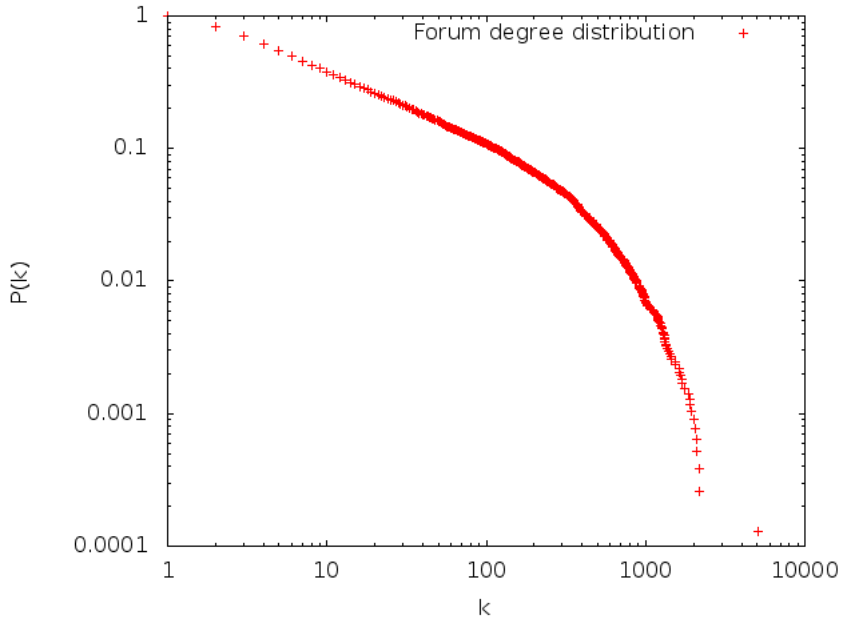


Figure 2.6: Forum-post degree distribution plotted with CDF method.

2.4 Clustering coefficient

Humans, like many other animals, need to interact with each other and make group in order to collaborate and utilize resources. Making ties between indi-

viduals give them more power and has many benefits for individuals as well. When it comes to networks, it is no surprise to see the same characteristic in human behavior. Human networks are known to be highly clustered [28].

A clustering coefficient tells how many triangles the network has compared to the maximal possible number of triangles (given the set of degrees of the nodes). Consider this very simple graph Figure 2.7:

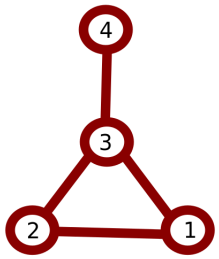


Figure 2.7: A simple graph with three closed triplets and two open triplets. Three closed triplets make one triangle.

By definition each three nodes connected by three edges, are called *closed triplet*. If one pair is not connected then it is *open triplet*. In the Figure 2.7, closed triplets are $[(1,2,3),(2,3,1),(3,1,2)]$ and open triplets are $[(4,3,2),(4,1,3)]$. The *clustering coefficient* is the number of closed triplet divided by all triplets either open or closed. See Equation 2.5. Three closed triplets make one *triangle*.

$$C = \frac{\#closed\,triplets}{\#all\,triplets\,(open\,or\,closed)} \tag{2.5}$$

Thus in this example, clustering coefficient equals to $C = \frac{3}{5}$. In section 4.10, we will discuss the results using above method.

2.5 Null model of a network

Like other physical systems whose state is defined against a reference frame, we need to define a proper reference for networks. A *Null model*, helps to define the network's structure—how the network differs from what would be expected by

the trivial properties (which can differ between systems). For example networks of humans are known to have higher number of clustering coefficient compared to null model, which means humans are more likely to make small groups in the network than expected by chance. Thus for our purpose which is focused on human networks, it is important to build a null model and compare parameters from a real model with a null model.

The idea of rewiring links as a null model for networks, first introduced by D. Gale in 1957 [8]. Their null model keeps the number of vertices with a certain degrees conserved and everything else random. Later on the null model improved in such that one selects two edges of the real model randomly, like edge (i, j) and edges (k, l), and shuffle the edges so that we will have new edges (i, k) and (j, l). However if one or both of this shuffled edges are already exists in the network, then one should select two other edges. This process is repeated at least for all edges in real model. The output is an instance of the null model for the real network [27].

We used this method of null model for most of the measurements in chapter 4.

2.6 Structural balance theory

The network we study has two important features: 1. It is temporal 2. It is multiplex. Here we will focus on the later aspect.

The community consists of two different communication environments. A Forum where people discuss movies or other movie-related topics, and one where each member can send and receive private messages from other members. These two media have different psychological and social values. As a result, we can expect people to act differently, and that this different behavior is reflected in the network structure. As a response to each discussion thread in the Forum, three things can happen: a member can read the thread and add a comment to a specific person, he can ignore the thread or he can send a message back to desired target. We assume each of these possibilities is associated with different psychological and social meaning. Another possibility is that two members only exchange private messages without making any comment in public Forum together[4].

Now let us imagine a group of three people who are interacting with each other. Psychologist Fritz Heider in 1946 built up a theory which explains the pattern of communications between three persons. In his theory which was soon generalized to *structural balance theory*, he explained how a social triads could happen in a society [30]. Structural balance theory explains the form of friendship between three people if they are connected together. The theory defines two concepts of friendship between people. If two people like each other, positive link. If they do not like each other as a negative link. Now imagine that you (A) are a friend with (B), and your friend (B), has an enemy (C). The

[4]This possibility is supposed to be low unless two people knew each other from before or other social reasons which we are not going through them.

theory suggests that to have a balance triad, it is more likely that you are also the enemy of (C). If this not happen, then the triad is not balanced which cause tension. Figure 2.8 shows different pattern of triads which can be balanced or unbalanced.

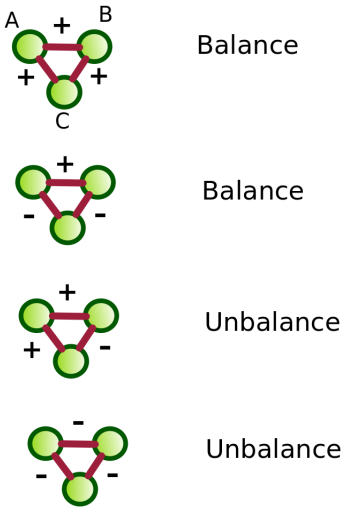


Figure 2.8: The graph explains structural balance theory. Links are either positive means "friend" and negative which means "enemy". If the links of three, makes a desire social contact with no tension, then it is balance. If the social triplets makes tension then it is unstable. First triangle shows the combination of "friend's of my friend is my friend."

In chapter 4 we will show the structure of the community compared with structural balance theory and we will discuss differences[5].

As mentioned before, the community does not only contain accessible Forum page for all members but it also has private space for sending messages between members. These public and private spaces will affect the pattern of communication and we believe that our network is not exactly following the same pattern as structural balance theory. We will also discuss how these private and public areas will affect the pattern of communication.

2.7 Assortative mixing in the network

In many real networks, nodes or vertices are connected by edges or ties which are not randomly happen in the networks. In the network of friendship for example the pattern of the network is affected by language, race or age between

[5]See section 4.10

friends. The network is assortative-mixing or assortative-matching if nodes in the networks tend to join with other nodes with similar characteristics. In social networks it is known as *homophily* [17]. If nodes or individuals tend to join with nodes with different characteristics, then the network is known as disassortative mixing. Thus the structure of many networks is influenced by assortativity mixing. Another example of strongly assortative mixing is sexual partnerships. In this network the ties are more common between men and women with common ancestry [19].

For measuring assortative mixing, first we define e_{ij} which is the fraction of edges which connect nodes of type i to nodes of type j in a network. Thus for an undirected network we have:

$$\sum_{ij} e_{ij} = 1, \sum_{j} e_{ij} = a_i, \sum_{i} e_{ij} = b_j \tag{2.6}$$

Where a_i is the fraction of edges which show connections of type i. By definition assortativity coefficient is:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{2.7}$$

When there is no assortative mixing $\sum_i e_{ii} = \sum_i a_i b_i$, therefore $r = 0$. In case of perfect assortative mixing, $\sum_i e_{ii} = 1$ and $\sum_i a_i b_i = 0$, which means $r = 1$. For disassortative network, $-1 \leq r < 0$.

Assortative mixing could also be calculated in terms of degree distribution, *assortative mixing by degree*. If vertices with high number of degrees are more likely to connect to neighbors with high number of degrees, then the network is assortative. If higher degree nodes are more likely connected to lower degree nodes in a network, then the network is disassortative. One should notice that assortative mixing by degree does not necessarily omply homophily, which we defined above. People can have similar characteristics but different number of connections depending on their social activities. For an undirected network, we have [18]:

$$r = \frac{4\langle k_1 k_2 \rangle - \langle k_1 + k_2 \rangle^2}{2\langle k_1^2 + k_2^2 \rangle - \langle k_1 + k_2 \rangle^2} \tag{2.8}$$

In the community that we study, there is no information about people in the community except a number which represents them. When two members send a post on Forum or send an email to each other then there is a link between the two people. In this case assortative mixing by degree is more suitable to calculate. See section 4.2.

2.8 Spearman correlation coefficient

Spearman correlation coefficient is used to show the statistical dependence of two variables. It can be varied from $+1$ to -1 depending on whether a monotonic

trend is increasing or decreasing. If two variables are completely monotonically related, then it will be ± 1 even though it is not linear relationship. This is the main difference to the more well-known "Pearson's correlation coefficient" that measures the difference to a linear relationship. Imagine we have n pairs of data (x_i and y_i) and we want to see the correlation between data in one variable with data in the other. Let's define $d_i = x_i - y_i$, the difference between two data observations. By definition, the Spearman correlation coefficient (ρ) is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.9}$$

Figure 2.9 shows a sample of scatter plot with corresponding Spearman coefficient. In the result chapter we will discuss and use Spearman coefficient for the scatter plots. See sections 4.10, 4.15 and 4.17.

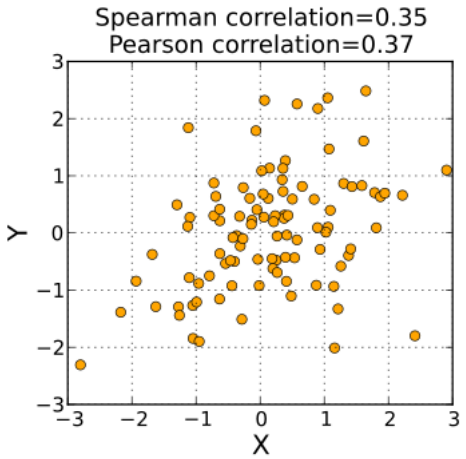


Figure 2.9: A sample of scatter plot. Spearman coefficient summarizes the correlation between the two variables in x- and y-axis. The image is taken from www.wikipedia.com.

Chapter 3

The data set

In this chapter we introduce the study case Filmtipset.se[1]. In section 3.2, we review one earlier work on a different data set from the same website.

3.1 Filmtipset.se at a glance

Filmtipset.se is the largest online community for discussing and reviewing movies in Sweden. It has been in operation from 2000 till now. According to the homepage, it has approximately 98000 members. In Filmtipset.se, you can get the latest news about movies, lists of currently screening movies, TVs or DVD. Filmtipset.se has a lot of minor features, for example, number of movie categories with top 50 movies in each. It makes it easier for the users to find their own favorites. It also has different general categories like best movies, worst movies, member's favorites and most rated movies. You can select your favorite category and see how other members rated movies and read about member's opinion about different movies and so on. Table 3.1 shows general facts about Filmtipset.se.

Figure 3.1: The header of the first page of Filmtipset.se

As a member, you have a personal profile in which you can see the statistics of your activities such as how much time you have spent in Filmtipset, the last

[1] To be clear in the text, we refer to Filmtipset.se by using first capital letter and sometimes we omit ".se" from the end. We also use "the community" refer to Filmtipset.

Number of members	97917
Number of movies in the data base	76868
Total number of ratings	21,399,647
Total forum posts	1602818
Dominant gender and age	Educated male between 25 to 34
Traffic by city	1. Stockholm 2. Lund 3. Uppsala

Table 3.1: Filmtipset.se in statistics. The last two statistics came from Alex Traffic Rank, www.alexa.com.

log-in time and so on. You can see who has visited your profile recently, you can send email to other members and receive emails in your mail box from other members. You can rate movies from very poor to excellent as a number from 1 to 5 which are also shown in different colors. You can also visit other member's profiles, the statistic of their activities, movies they have rated and you can add them to your friend's list[2]. Figure 3.2 shows a sample of a member's profile who happens to be the advisor of the Thesis.

[2]This is not like Facebook.com where the other part has to reciprocate this for a "friendship link" to be established. It only means that you add a link to that person's homepage so that you easily can see your friends latest film grades etc.

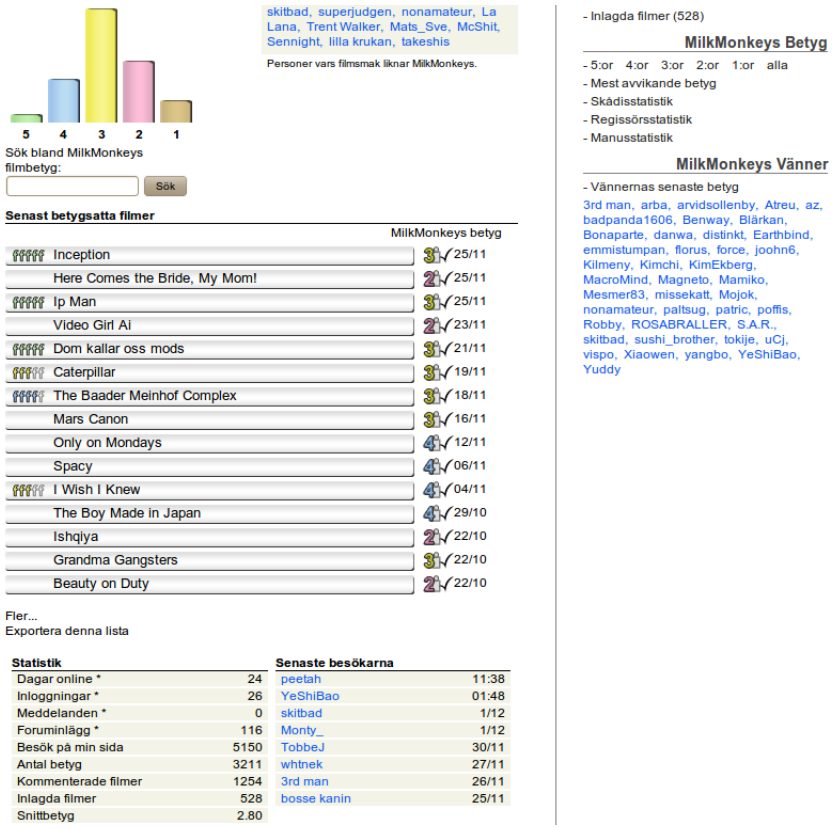


Figure 3.2: A snapshot of member's profile in Filmtipset

One important feature in Filmtipset is forum page. In this section, members can open new discussions or contribute thread of discussions. Members can directly post comments to whom they want to reply so that the discussions become like trees. The form of threading helps to keep track of who has sent a message to whom. Figure 3.3 shows how the forum page looks like.

Figure 3.3: A snapshot of a Forum thread in Filmtipset

This project is based on the data sets from forum posts and private messages between members. The data set contains an id number which corresponds to each member, and a time when a message was posted by a member to another member either in forum posts or by sending email[3]. The data sets do not contain the details of the messaging between members, they only contain user id and time.

3.2 Early study about Filmtipset.se

There has been one study of Filmtipset before by a group of people from Technische Universität Berlin [25]. In their paper they focused on two different kinds of relationship between members: friends and fans. The authors define *Friends* as reciprocal relationship which means both members add each others to their friend's lists. But if only one of them add the other member to his/her friends' list and not vice versa, the relationship is called *fans*. The authors showed the correlation of degree distribution of friends and fans in the network[4]. They also showed the relation between the number of friends someone has and the number of movies has been rated by him/her. They claime that if a member has large number of friends, he/she will often has rated a large number of movies. In other words, this member is an active member in Filmtipset. Another aspect in the previous study of Filmtipset is measuring Jaccard similarity coefficient in terms of similarity of two pairs (two members) if they have watched the same movie or not. They used different scenarios for comparing similarities. Their results suggest that there is a significant correlation between friendship and movie taste[5].

[3]The data is anonymous so that we cannot match a data entry to a username

[4]Similarly we measured reciprocity in the network. See section 4.3

[5]we use Jaccard similarity method in chapter 4, but my focus is similarity between member's connections than movie taste. See section 4.11

3.3 How does the community look like in network perspective

Figure 3.4 shows a small section of how the network of interaction between members in Filmtipset.se looks like. Each circle shows a member of the community. Members are interacting with each other in different ways. They can send private messages to each other shown by green links in the figure. Members can open a new thread of discussion or comment to other comments in public Forum page. These multiple interactions give us a chance to look at a more realistic social network. Such networks are known as *multiplex* or *multi-relational* network. Furthermore, we know when the contacts along an edge has taken place. Each link can be associated with list containing the times when message or comment has been sent. This time list, helps us to understand and study the time evolution of the community and also time evolution of interactions between members. For example, we can investigate how long it takes for people to answer the emails or what pattern of human activity looks like. Another aspect of the community that helps the data analysis, is that it is a closed society without interaction from the outside. This is in contrast to many datasets about human communication studied in the literature where only a part of the nodes are sampled. In such cases one gets difficult sampling problems about how to handle the boundary between the sampled nodes and the rest.

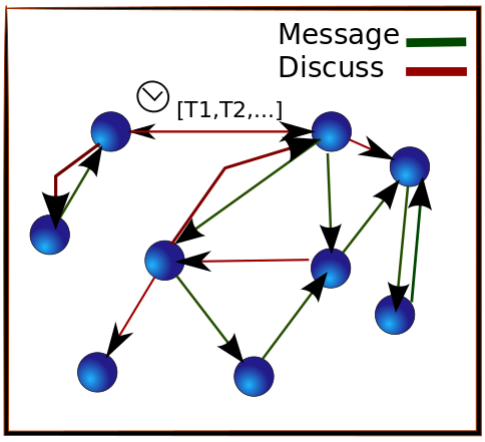


Figure 3.4: Pattern of interaction between members in the community. Green links show sending messages and red lines show sending comments in Forum page. Each link contains an attribute, a list of time which shows when a message (or Forum comments) has been sent. By this attribution we avoid multiple links. The community is enclosed and there is no interaction from outside. One needs to register at the website to be able to interact inside the community.

Chapter 4

Results

Introduction

In this chapter we discuss results from studying the online movie community[1]. We used measurements methods were introduced earlier in chapter 2. Here the network consists of members of the community as nodes in the network and each edge represents a message or a post which has been sent from one member to another in Forum page or private Messages. In some cases we consider the network or graph as directed graph and in some cases undirected graph[2].

4.1 The community over the years

We start by giving a general overview of the time evolution of the community, to see how a community starts and develops over time. Here we consider the total number of posts in Forum and the number of Messages sent by members for each month as a *popularity* of Filmtipset. The time duration of the data set is 7 years. We measured the number of posts in Forum and number of messages sent as Messages separately.

[1] In the text, whenever we talk about 'the community', it means Filmtipset.se

[2] For the sake of simplicity, when we use *Messages* with a capital letter, that means the data set corresponding to private messaging between members. When we use *Forum* with the first capital letter , that means the data set corresponding to member's posts in Forum pages.

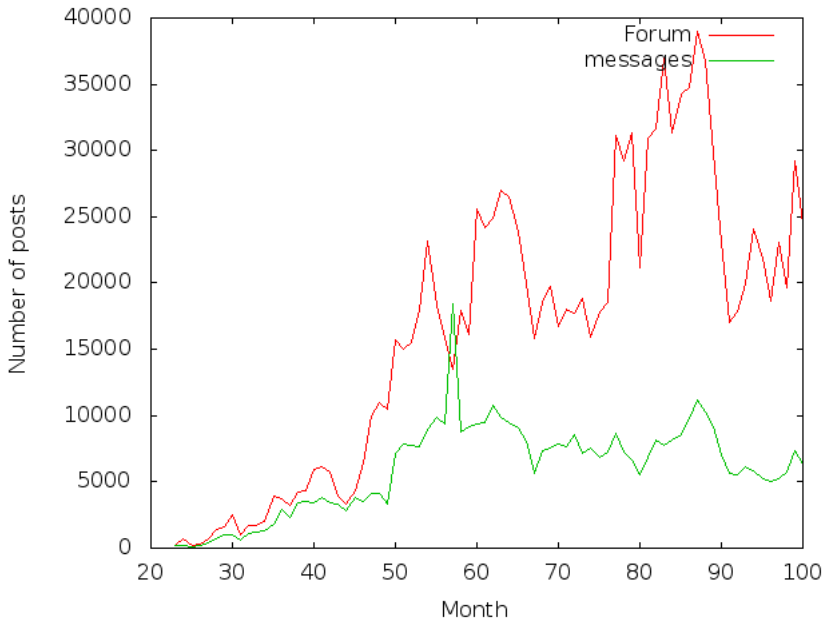


Figure 4.1: Number of posts and messages sent by members to Filmtipst in every month during 7 years.

Figure 4.1 shows people's activity in each month. The activity of posting in Forums or Messages are almost following the same pattern. The activity in Forum posting has more peaks which relates to popular or important topics which pop up.

4.2 Assortative-mixing pattern

Here we measure assortative mixing by degree. The aim of this measurement is to see how people are connected depending on their degree. Is it more likely that people who have many connections to other people (high degree), are connected with each other (*assortative*)? Or do people with high number of connections more likely connect to people with a lower number of connections (*disassortative*)?

Table 4.1 shows assortative-mixing coefficient in Forum and Messages networks. To be able to see the significance of the pattern, we compare the results for the empirical networks with null models (rewired-networks) for both cases. For the null model, we took 10 samples and take an average of the assortativity and we measured the standard error.

	Real model	Null model	Standard error for null model
Forum network	−0.17476	−0.17979	0.00012
Messages network	−0.03774	−0.04403	0.00026

Table 4.1: Assortativity-mixing coefficient for Forum and Messages. We compare results with null model by randomizing the edges but keeping the number of nodes and their degrees.

In social networks the assortative-mixing coefficient is known to commonly have positive value, which means social networks are usually assortative [19]. However our results show negative value for assortative mixing. This online community is disassortative compare to random model which gives $r = 0$. In other words, in this community members with high number of connections are more likely connected to people with low number of connections. Another interesting observation here is that results do not show a significant difference between the assortative coefficient in real model and null model which looks like neural network [12]. This slight difference between assortativity of the real model and rewired network, tells us something about neutrality of the network. To see the tendency of shifting of assortativity coefficient from null model we use *z-score* or *standard score*. By definition:

$$z = \frac{\{\text{real-null}\}}{\sigma} \tag{4.1}$$

Where σ is standard deviation. Standard deviation is multiplying standard error from the above table to square root of number of samples, in this case 10.

Thus the Z-score measures the deviation from the expected value in units of the expected standard deviation. Z-score of assortativity for Forum network is 13 and for Messages is 8. This means that in Forum there is more tendency to be more assortative than in messages compare to null model.

4.3 Reciprocity in the community

Reciprocity, as we introduced earlier, tells how probable it is in the network that edges or links between pairs are mutual (in both directions). In friendship networks, links means friendship between two people. In this network however, links have a different meaning. In this community, links show sending a private message or sending a comment to someone in public Forum. Thus, direction of links show who is sender and who is receiver. Mutual links show if a pair send back to each other or not. Results show reciprocal connections both in Forum and Messages networks. See Table 4.2.

	public Forum	Private Messages
Reciprocity	0.59	0.65

Table 4.2: The table explains reciprocity in public Forum and private Messages. In both Forum and Messages reciprocity is positive and high.

From the Table 4.2, one could explain why reciprocity in Forum and Messages is positive. In this community people care to follow and send back a comment and message. It is more likely that someone who gets a private message, will answer the message, while there is less pressure to answer back public comments which is clear from the results.

4.4 First contact: Where do people start their first connections? Private Messaging or common Forum posts?

we are interested to see how two people start contacting each other and initiating an online relationship. Do they prefer to start the connection in public places or private places? Is there any difference between off line communities[3] and online communities[4] in terms of starting a connection?

Our sample is special and might not completely generalize to the entire society. In Filmtipset unlike other online social networks, like Facebook or dating websites, most members have at least one same interest in common: they like movies. On one hand it might be easier to start sending private messages relates to their movie tastes. Then yet again, people may feel that there is less of a pressure to build a relationship in off-line communities compare to off-line society. In the latter, one could easily control his/her privacy space. However our expectation from the data set is that, although it is possible to start a relationship from sending private messages, members tend to get to know each other from Forum pages that is public place. It is easier to start new contact in bublic than just hit the privacy bubble all of a sudden[5].

To have better sample, we eliminate pairs of people who only exist in one data set (only private messaging or public Forum). By this elimination, we make sure to look at only pairs who have connections in both private and public place. For the simplicity, we imagine the graph is undirected; therefore, it doesn't matter who is sender and who is receiver of the text. In table 4.3, we measured the proportion of people who start the first contact from Forum posts and from private messages. To be able to see if the result is significant or only come from the structure of network, we build a null model and compare our result with

[3] like society which is not online

[4] like online social networks

[5] One online article with the same expression, explains why French people seem to have privacy bubble for new comers although in 'Latin culture' close physical distance between two people is acceptable. The bottom line of the article is that French people strongly respect other's privacy and they give a new comer time to find his or her space in a relationship [1]

null model. Our null model consists of same people who exist in the real model, but we randomize the connection times. With this change, the first connection is just randomly distributed. Table 4.3 shows results for the real model and the null model.

	Percentage of members (pairs) start first contact in Forum	Percentage of members (pairs) start first contact in Messages
Real model	67%	33%
Null model	47%	48%

Table 4.3: The percentage of pairs who start their first contact in Forum and messages compare with null model

4.5 Degree distribution

As we mentioned earlier, degree distribution is an important signature of the network. It helps to understand if people make contacts by random chance or by some active considerations.

From a network point of view, degree distribution gives information about how diverse the number of neighbors of vertices are. We use different methods of measurements to look at degree distributions; First we will discuss the theoretical background.

4.5.1 Cumulative Distribution Function (CDF) method

As we mentioned in section 2.3.3, the Cumulative distribution function (CDF), is an useful method for plotting degree distribution. We use the CDF to visualize degree distributions. Except for the undirected degree, we also plot in-degree and out-degree distributions. In this case in-degree represents number of messages which are sent towards somebody in public Forum or private Messaging and out-degree represents number of messages which are send out by members. Figure 4.2 and 4.3 show CDF plots of the Forum and Messages data.

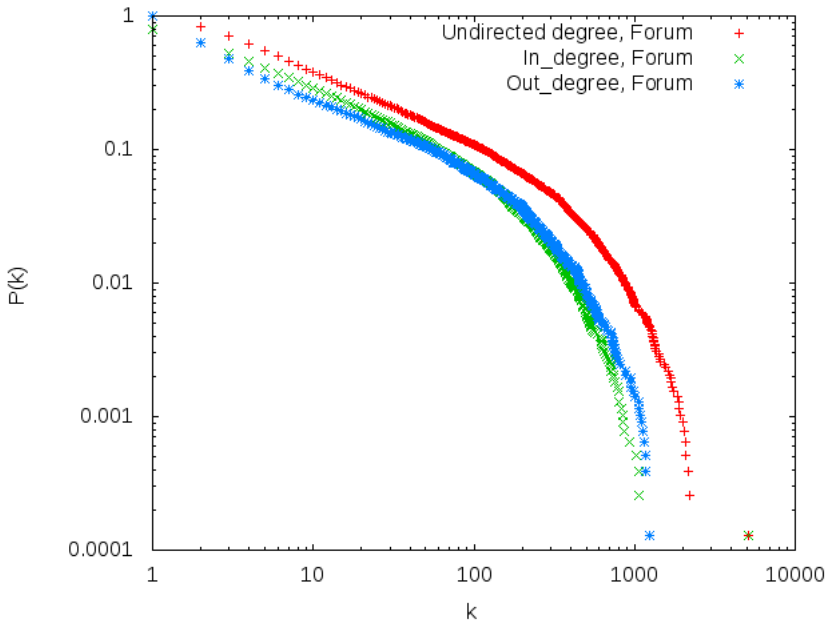


Figure 4.2: Cumulative distribution function for the degree of posting a message in public Forum. The blue line shows out-degree and green line shows in-degree distribution. The red line shows general degree distribution. Figure is in logarithmic scale.

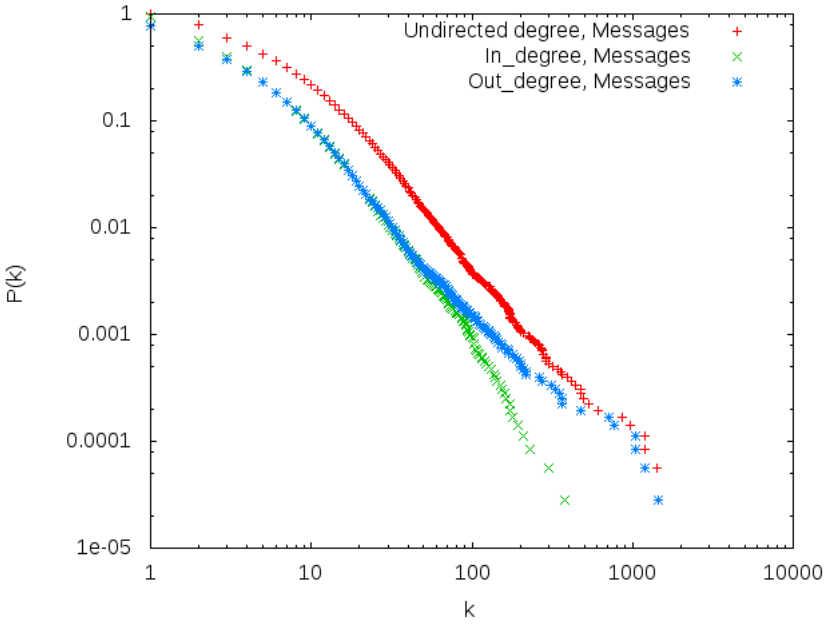


Figure 4.3: Cumulative distribution function for the degree of posting a message in private Messaging. The blue line shows out-degree and the green line shows in-degree distributions. The red line shows undirected degree distribution. Figure is in logarithmic scale.

4.5.2 Fitting to power law

From the figures shown in previous section, one can notice that the degree distributions are following power law to some extent. However, they seem to follow a power law, not completely but with a cut off. Thus we fit Forum and Messages degree distributions which we got from CDF method with a power-law-with-a-cut-off equation:

$$y(x) = ax^{(-b)}exp(-x/c) \tag{4.2}$$

Where $exp(-x/c)$ is the cut off. Fitting Forum and Messages degree distributions with power law with cut off, show that for Forum posts, it fits nicely with the graph. The values of b and c from non-linear curve fitting are 0.4892(8) and 654(6) respectively (the numbers in parentheses indicate the standard error of the last decimal). See Figure 4.4, 4.5 and 4.6.

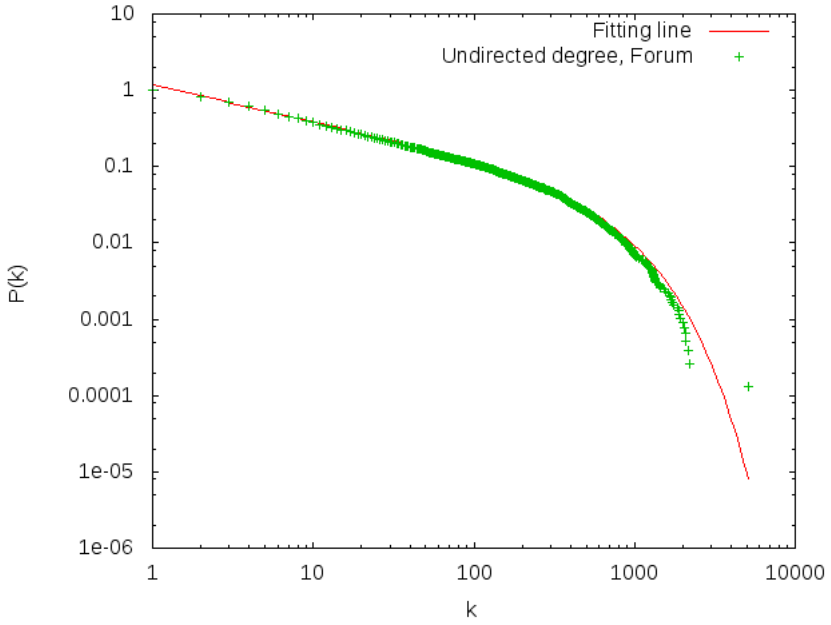


Figure 4.4: Fitting Forum degree distribution with power law with cut off. The green line is the degree distribution and the red line is the fitting line. The plot is in logarithmic scale.

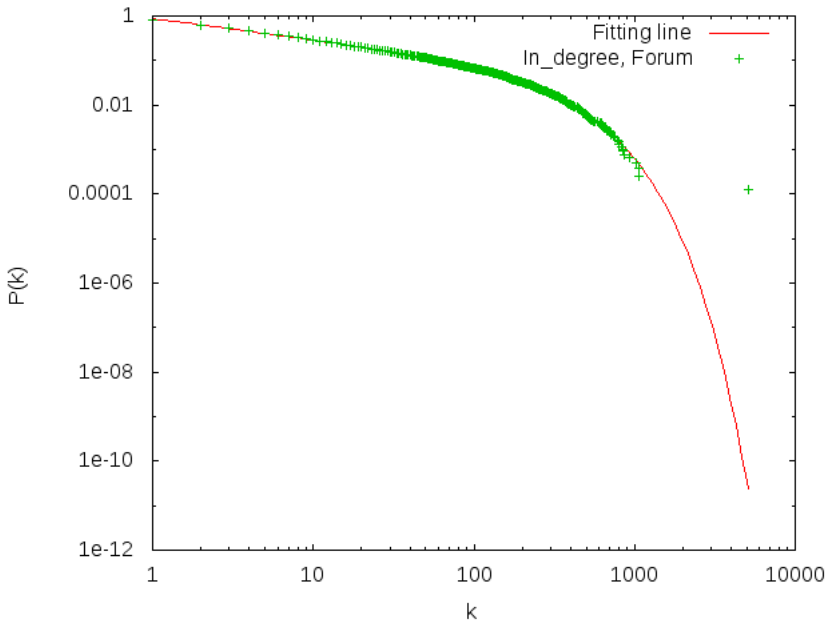


Figure 4.5: Fitting Forum in-degree distribution with power law with cut off. The green line is the degree distribution and the red line is the fitting line. The plot is in logarithmic scale.

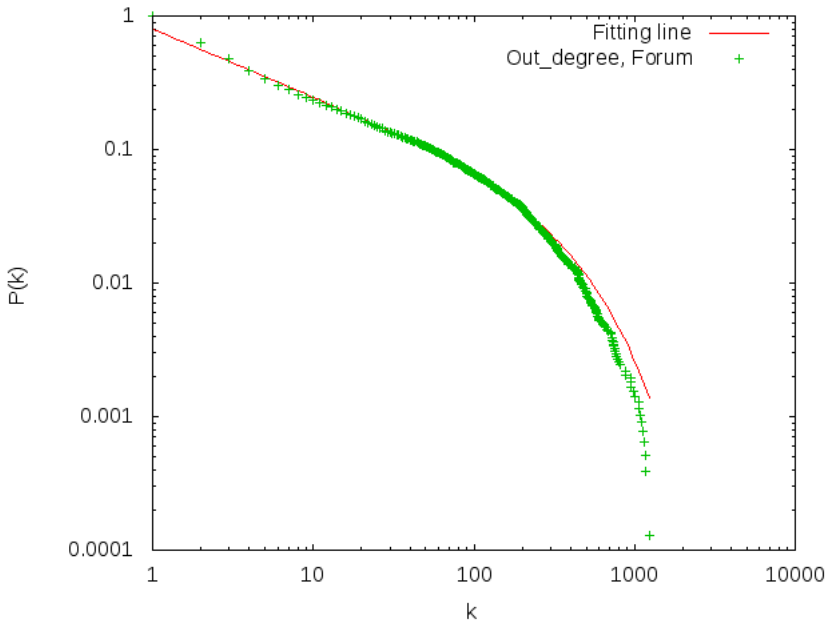


Figure 4.6: Fitting Forum out-degree distribution with power law with cut off. The green line is the out-degree distribution and the red line is the fitting line. The plot is in logarithmic scale.

For messages, the fitting shows that the degree distribution does not follow any kind of power law. This kind of behavior is also observed recently in another study [29]. They claimed that ties with positive value make a degree distribution which does not have power-law behavior. In our case we can also consider links in messages as positive links because members who are sending messages to each other are mostly friends[6].

4.5.3 Logbin representation of degree distributions

We used logbin method to present the degree distributions and compare different networks with each other. See Figure 4.7, 4.8 and 4.9.

[6]Balance theory also suggests friendship attitude in Messages data set, due to over representing of Messages triangles in the community. See section 4.10.

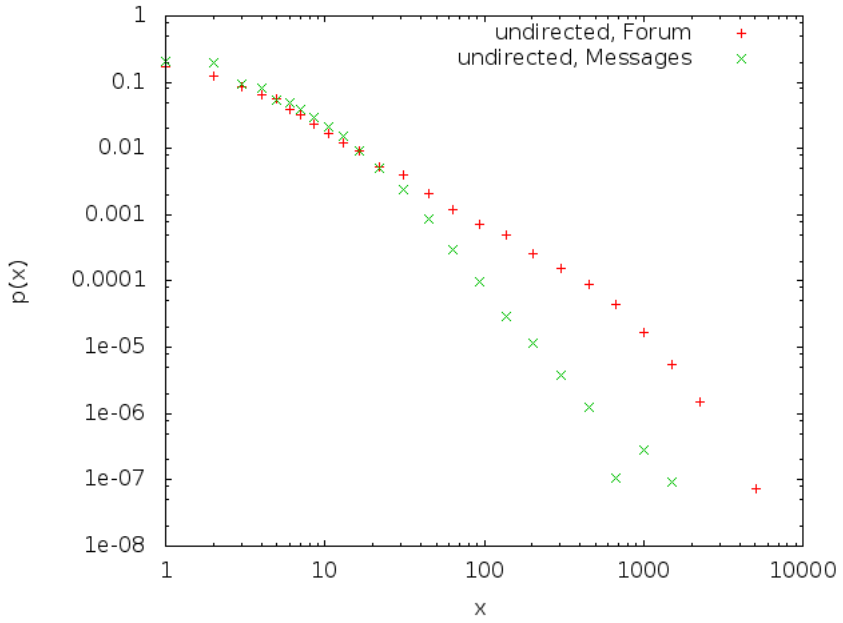


Figure 4.7: Degree distribution plotted by the logbin method. Red dots correspond to undirected Forum and green dots to undirected Messages data set.

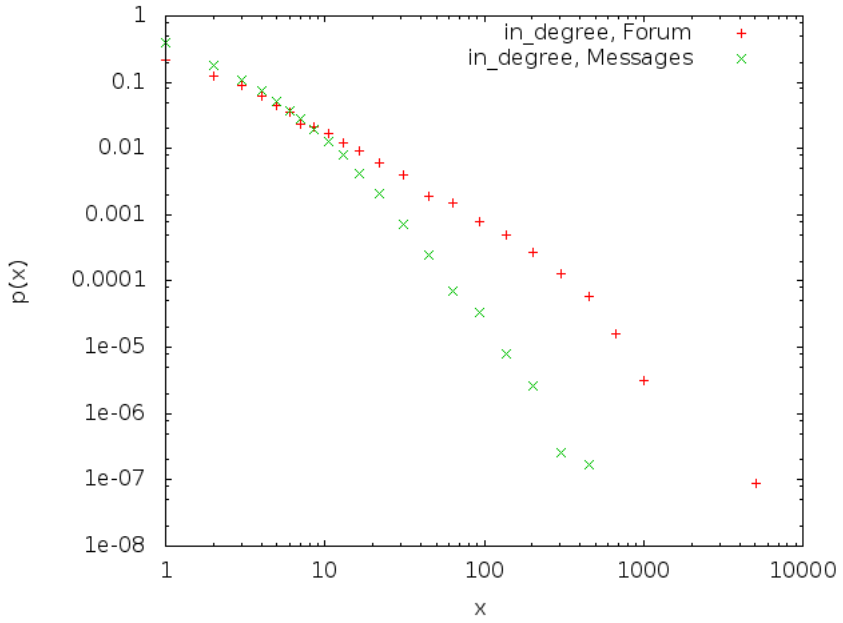


Figure 4.8: In-degree distribution plotted by the logbin method. Red dots correspond to Forum and green dots to Messages data set.

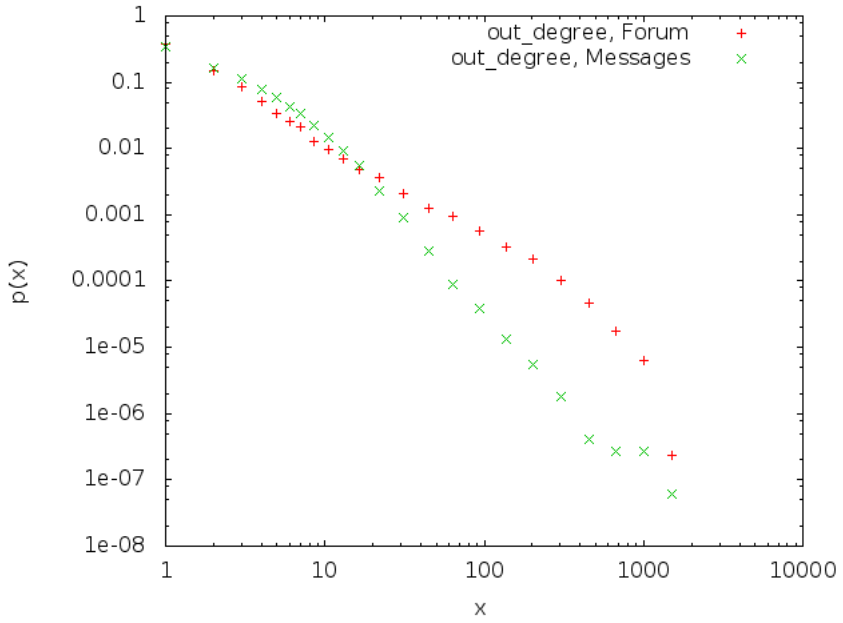


Figure 4.9: Out-degree distribution plotted by the logbin method. Red dots corresponds to Forum and green dots to Messages data set.

We can also Compare degree distributions for each member in Forum and Messages network. Here we measure number of connections – number of neighbors for each member– comparing Forum posts and Messages. We want to know if people who are actively posts in Forum to different people, are they also have many contacts in their email Messages. To test this hypothesis, we plot the number of connections for each member in Forum and Messages. Plot shows the correlation of connections in Forum vs. Fessages. See Figure 4.10.

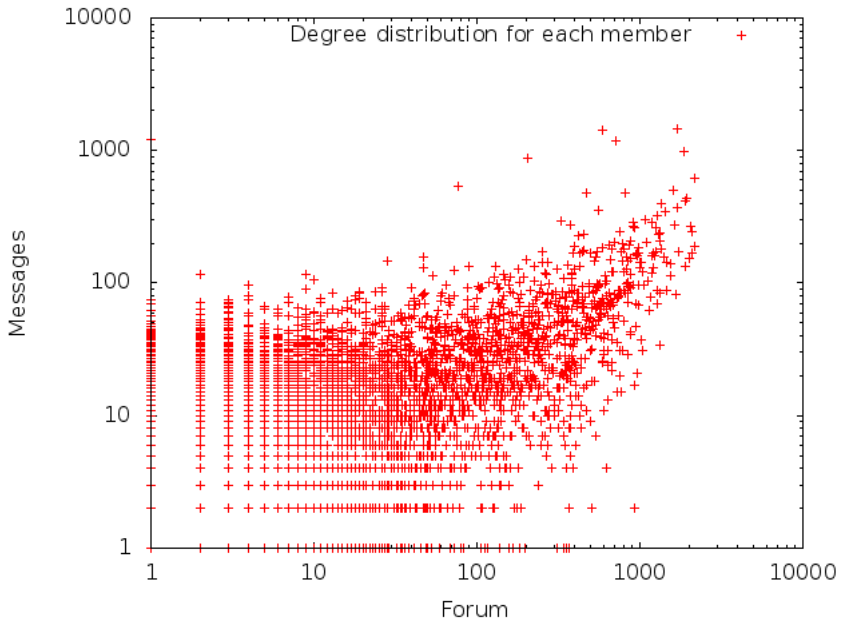


Figure 4.10: The scatter plot shows degree distribution of Forum in x-axis and Messages in y-axis for each member in the community network.

To understand the scatter plot in 4.10, one can condense it down further to just one number quantifying the correlation. One method to see if there is a significant trend in the plot is to use Spearman correlation coefficient[7]. In this case the Spearman coefficient equals to 0.15 which is positively correlated. Result shows it is most probable that members who are actively sending or receiving Forum posts are also active in private Messaging.

4.6 Response time and daily routine

A comment is posted in Forum page or an email send to a member, the question is how long does it take for the receiver to respond. Is it any difference in the response time in Forum posts and Messages or are they both following the same pattern?

Our experimental setup consists of a directed graph (Forum and Messages separately) and members as nodes in the graph. If there is any contact between two nodes then there will be an edge which consists of a list of time while each of the pair send a post (in Forum) or a messages (in mailbox) to the other pair. Then we go though the time list for each pair and measure the duration time between a post and a reply to the post.

[7] See section 2.8.

Figure 4.11 shows the pattern of response time in both Forum data set and Messages in logarithmic base. We used logbin method for plotting the results. One interesting result is the first peak in the pattern which corresponds to 14-15 minutes. Which means it is more probable that someone response to the posts or mail during 15 minutes after receiving the message. There is also another small discontinuity at around 1 day, which more or less shows that after one day of receiving a message, the probability of responding to that message decreasing more dramatically.

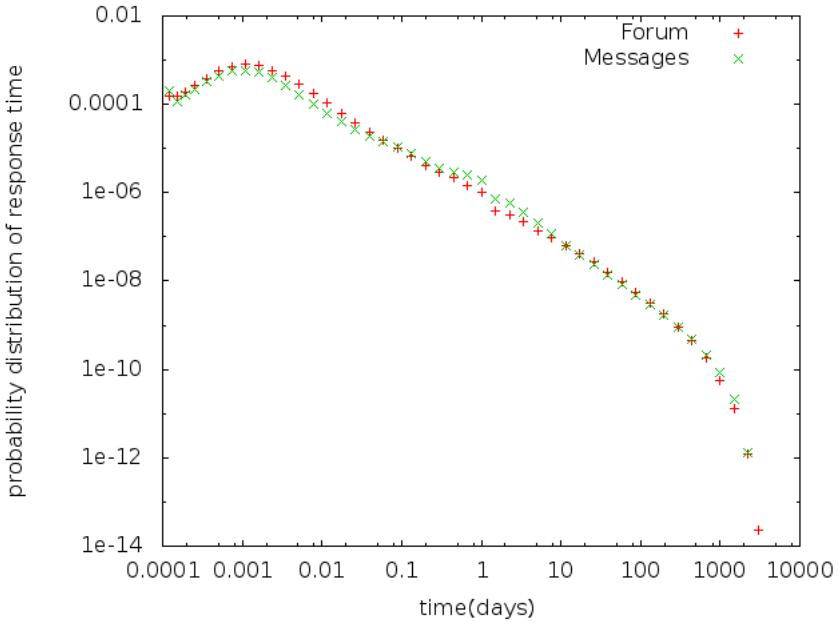


Figure 4.11: Response-time distribution, comparing Forum and messages data base. Plot is in logarithmic scale.

We can also measure histogram of the response time over days or weeks. It tells us the daily routine of members in terms of responding and replying to their comments and messages. We measure the proportion of number of responses per hour over all the responses for two time windows: two days and two weeks respectively. See figure 4.12 . Except for the first hour after receiving posts that Forum response time is significantly higher than Messages, they follow the same pattern in Forum and Messages for other hours.

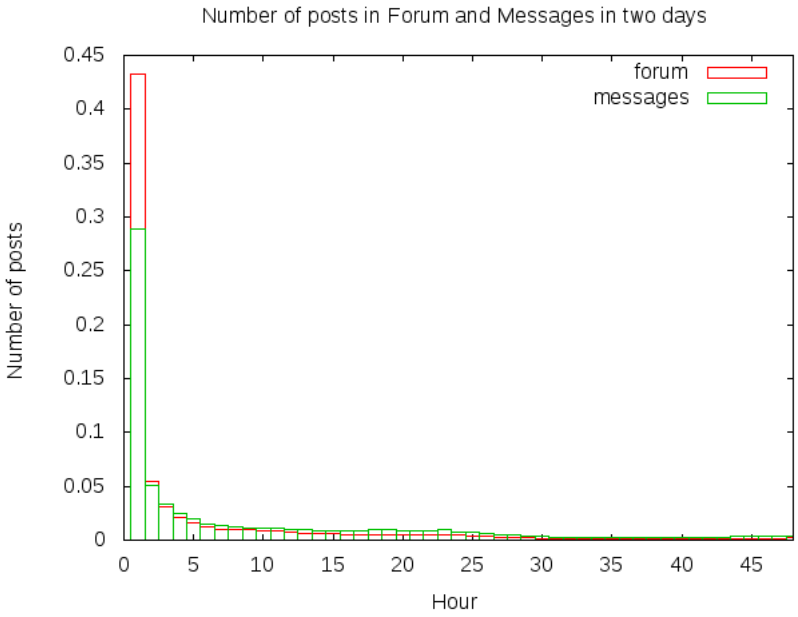


Figure 4.12: Histogram of response times for Forum and Messages in the time-window of two days. Except for the first hour that responses in Forum is higher than Messages, Forum and Messages are following the same pattern.

Figure 4.13 shows a periodic pattern which comes from daily patter of human activities. It shows when people log-in in the community and response to their messages. One can notice peaks in every 24 hours which represent daily routine. Responses in the Forum are more highlighted in the first day of activity, however responses in Messages could last longer. This tells time value of information spreading in the network. In Forum posts, it is more important to response early, but in Messages the response time could have more delay.

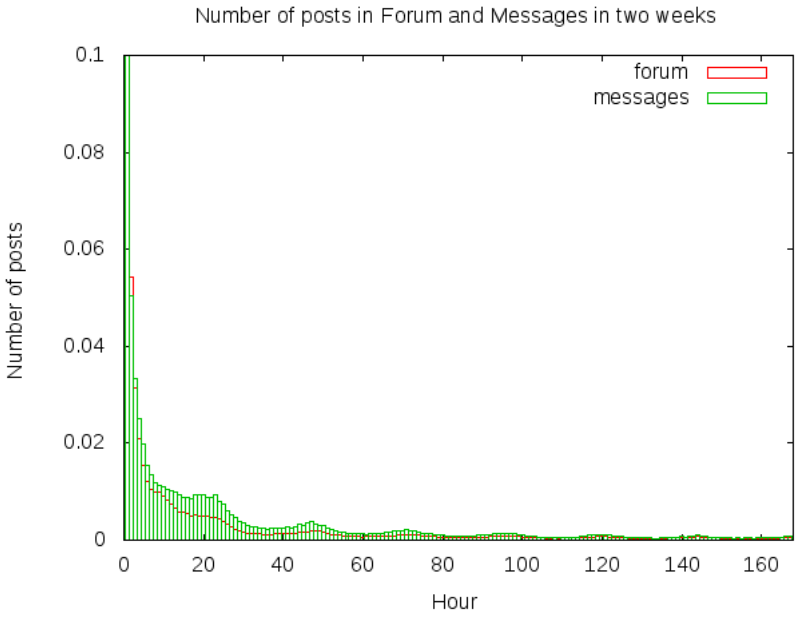


Figure 4.13: Histogram of response times for Forum and Messages in the time-window of two weeks. One can see the periodic pattern of human activities.

4.7 Member activity in the community

First of all we need to define what does activity mean in this context. The *activity* means how many members someone has contact with during his/her membership lifetime. Technically speaking, we measure number of out-degrees for each node divided by time interval between the first time the node appeared in the data set until the end time of data set which is the same for all nodes. Therefore, we imagine that all members are exists until the last date of data set which might not be true for all cases but it is good enough for our purpose. Figure 4.14 shows the activity of members in Forum posts and Messages by the help of CDF method.

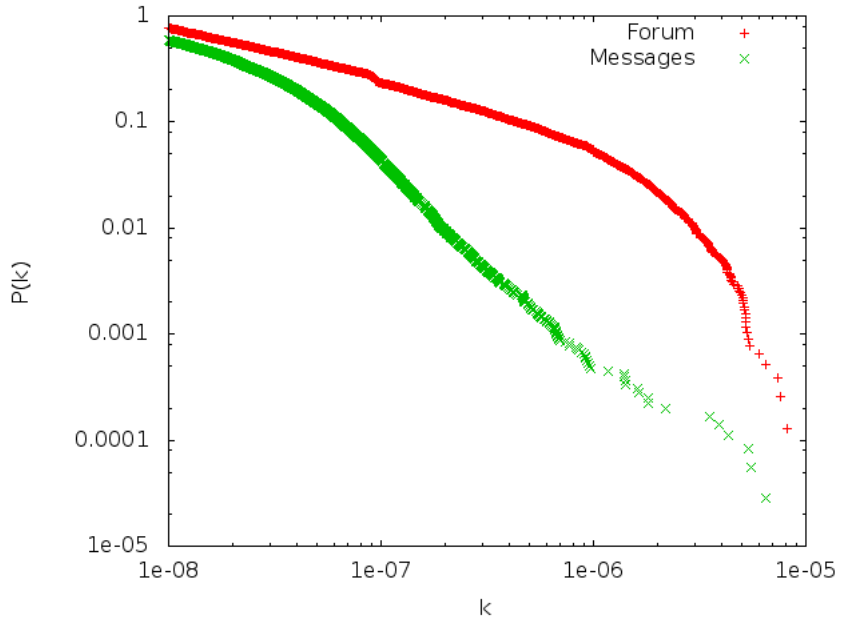


Figure 4.14: Activity as number of contacts for members in their membership life time, comparing Forum and Messages data set. Figure is plotted by CDF method.

Figure 4.15, shows the correlation between the activity of each member in Forum vs. Messages. The Spearman correlation coefficient in this case is 0.3844 which shows the increasing trend of activity in both axis. Which means members who actively send posts on the Forum to different people are also sending messages to many other members during their membership life time.

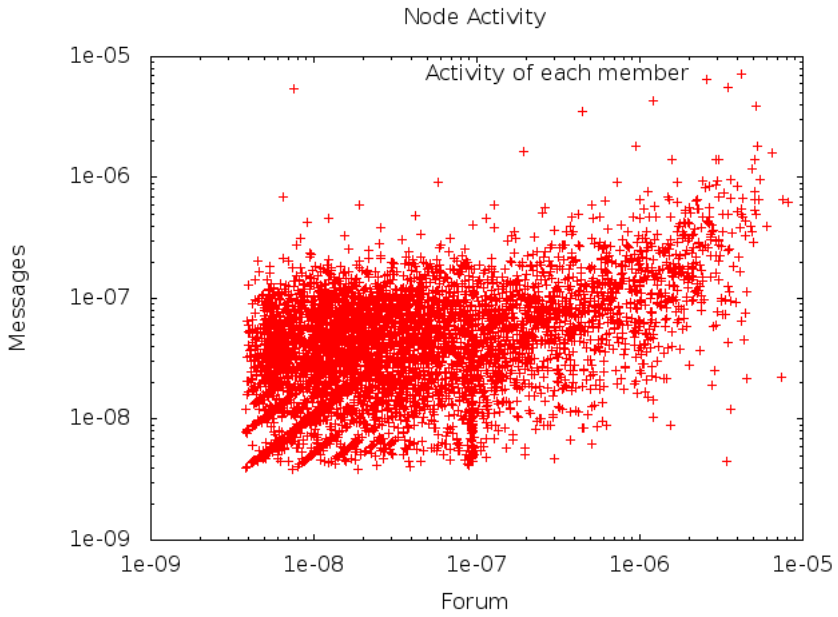


Figure 4.15: Scatter plot of activity as number of contacts for each member in his/her membership life time, comparing Forum posts and messages data set.

In the next step we define activity of a member in the lifetime as number of Forum posts or Messages one member has sent to the community in his/her membership lifetime. How many posts one wrote in Forum pages and how many Messages one sent to other members during the membership time. In Figure 4.16 we used CDF to plot activity of members per lifetime for Forum and Messages.

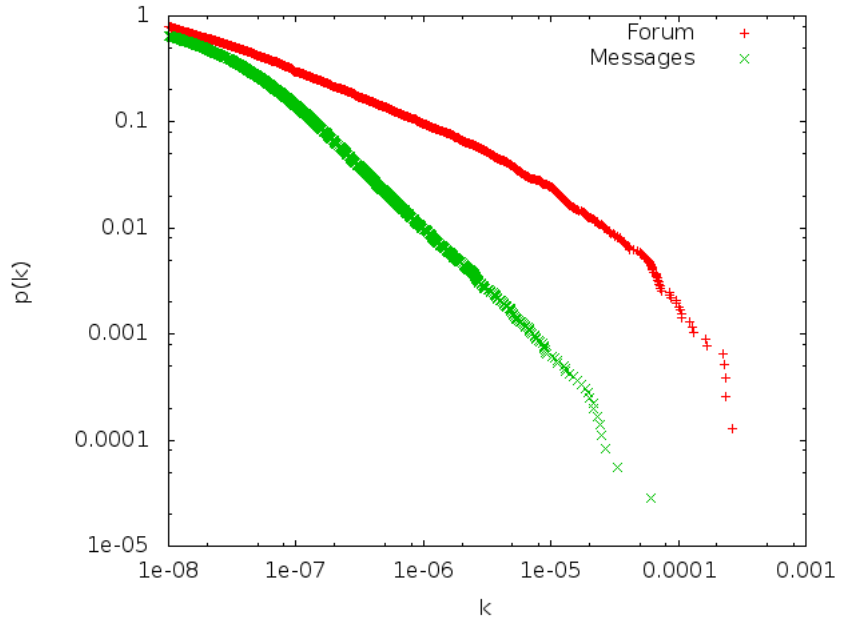


Figure 4.16: Activity of members as number of posts and messages in their membership life time in the community.

4.8 The activity status between two members

Imagine two members of the community: Alice and Bob[8]. Alice can send private Messages to Bob, or she may post a comment to Forum page as a reply to Bob's Forum posts. The question is how many private messages Alice sends to Bob compare to public Forum posts? Is the chance of sending private Messages or public Forum posts are the same for all pairs? Or is it the case that members have different categories of people to contact. For example if Alice and Bob are close friends they might prefer sending private messages instead of discussing in public Forum.

In the network point of view, we have two directed graph correspond to Forum posts and Messages. We look at each node and count how many edges it has with other nodes. Each edge is associated with two types of information: 1. The edge attribute which tells if the edge between two nodes are due to Forum posts or Messages or both. 2. A list of times corresponding to the time of sending Messages or Forum posts. Technically speaking, we measure the length of the time list for each pair of nodes for a directed graph.

[8] Alice and Bob are the main characters introduced by computer scientists to explain quantum transportation! Although it would be more appropriate if we would choose Swedish names!

The Figure 4.17 shows how many Forum posts and Messages sent by each pairs of members.

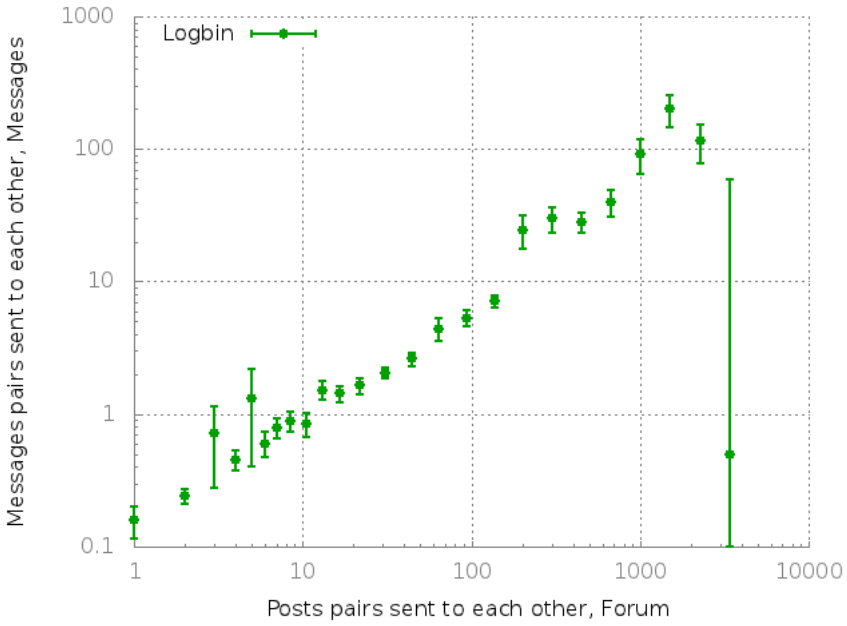


Figure 4.17: Logbin plot shows how many posts in Forum and how many messages have been sent by two members to each other in the community.

4.9 Interevent time

Time duration to send a post in Forum or a message in Messages for each member is called *Interevent time*. It shows the routine of member's activity and shows the time interval of sending a message in the community by each member [10].

For measurement, we consider a network with members as nodes. No matter who are receivers of a contacts, we only consider the time when a node sends out a message or post a comment in Forum posts. To measure interevent time we calculate the time duration between two following activity by each node or member.

The Figure 4.18 shows the interevent time for members in the community, no matter sending private messages or posting in Forum pages. We used logbin method for plotting. The figure is in good agreement with the results from the response time activity, See Figure 4.11.

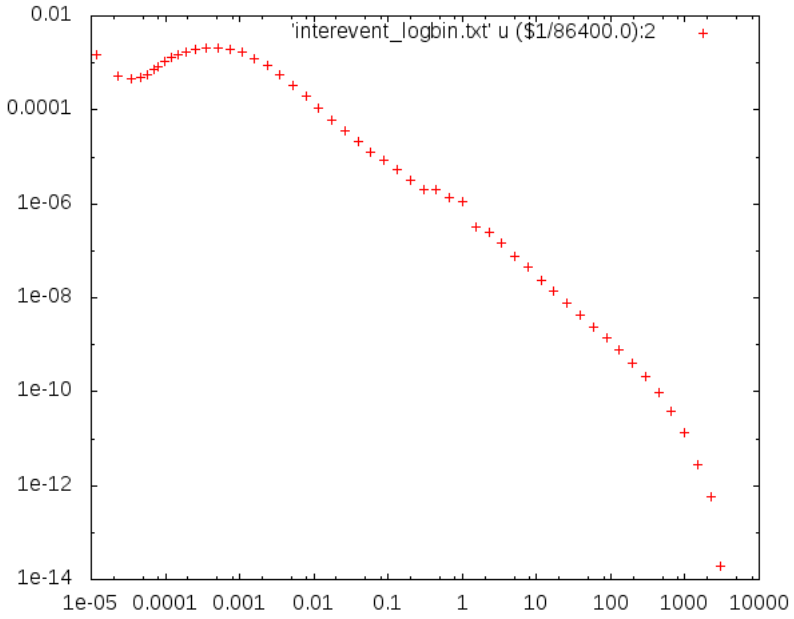


Figure 4.18: The interevent time of members in the community. We used logarithmic binning for plotting.

In the next step we consider interevent time for Forum and Messages for each member separately. The aim is to see if there is any difference in the pattern of activity in terms of time and information value. See Figure 4.19. From the figure, One can notice the first peak in messages happens earlier than Forum posts. After one day, both Forum and Messages interevent times, are decreasing, but in Messages, the decreasing rate is slower than Forum. This difference in decreasing trend suggests that participating in Forum pages needs shorter time interval however in Messages, delay in replying messages is reasonable.

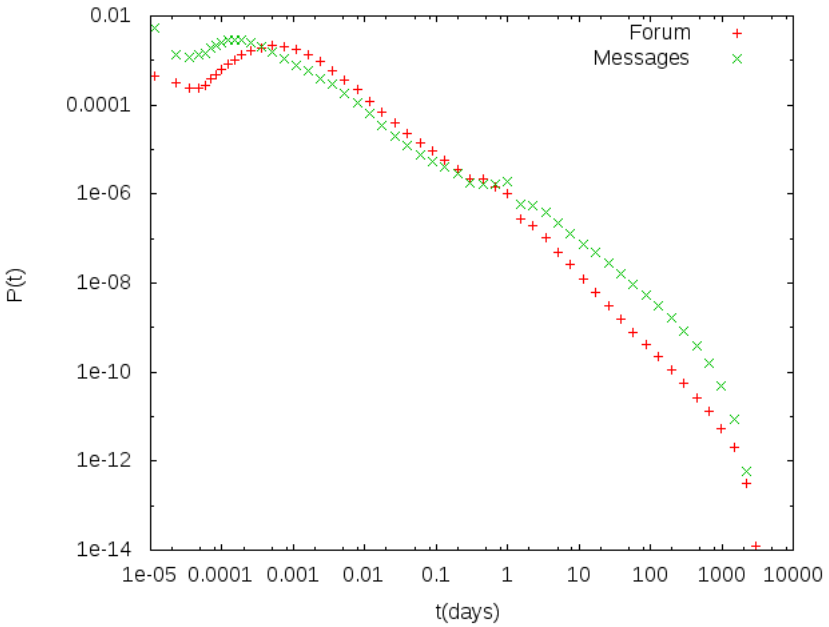


Figure 4.19: The interevent time of members in the community, depending on posting in Forum or sending private Messages. We used logarithmic binning for plotting.

Figure 4.20 shows the histogram of interevent in each hour. One could notice the member's routine of visiting and sending message in the community. The figure shows increasing trend about 8 to 10 hours after the first time, which relates to the fact that people are sending a post or messages once during work hour once after work or in the evening.

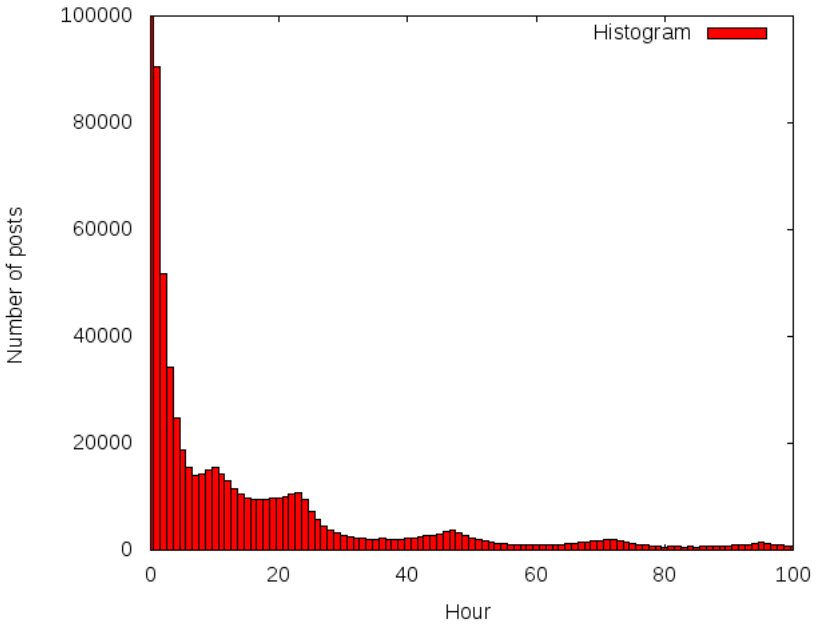


Figure 4.20: The histogram of interevent time of members in the community for each hour.

4.10 Triangles and structural balance

In this section we measure number of triangles in the network. Consider a small part of the network as we show in Figure 4.21. The red edges representing the connection via Forum and green edge a Messages connection. From this subgraph one can see a triangle between nodes 1,2 and 3. What we want to measure is this triangles in the network. Measuring the whole triangles divided by number of nodes get average clustering coefficient. However the goal is not to measure clustering coefficient, but to test the hypothesis (assuming private messages are, on average, more positive than the public ones) that there is a tendency for structural balance.

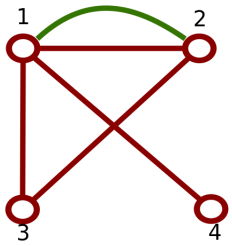


Figure 4.21: A sample shows part of a graph with two different edges. Red edges show postings in Forum and green edge shows sending messages in Messages network. Two triangles can be seen in this graph, one which has nodes 1,2 and 3 we call it FFF triangles and one which has the same nodes but with an edge from Messages we call it FFM.

Going back to the example, we see two types of triangles. If the red edge is called 'F' and green edge is called 'M', in this example, there is one triangle of 'FFF' and one triangle of 'FFM'. In reality it means that three members are connected together in the community (We consider undirected gragh here). Person 1 is connected to person 2 via Forum page and also by Messages (they comment to each other in Forum posts and they send a private messages to each other). Person 1 is connected to person 3 only via Forum. Person 2 and person 3 are also connected by Forum.

Recall the structural balance theory[9], one could label the edges as positive $(+)$ or friendship and negative $(-)$ as enemy relationship. Consider the relationships between people in a friendship network, triangles of people are more likely to be: $(+++)$, $(--+)$ and NOT $(-++)$ or $(---)$. Here we use the same concept to check if we can see the structural balance in the network. In this case for instance we imagine Forum edges as negative links $(-)$ and Messages edge as positive links $(+)$. The question is: Does the community network follow the structural balance like as real friendship networks?

We measured number of triangles in the whole network in the community. We consider triangles with 4 different patterns, triangles of:

1. Forum Forum Forum (FFF)

[9]See section 2.6

2. Messages-Messages-Messages (MMM)
3. Forum-Messages-Forum (FFM)
4. Messages-Messages-Forum (MMF)

Table 4.4 shows numbers of each kind of triangles in the network. Once again we need to compare our model with a null model to see if the results are significant or they are just because of the characteristic of the network. Our null model here is generated by shuffling the signs of edges between nodes.

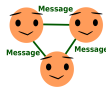
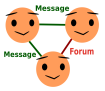
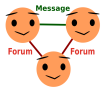
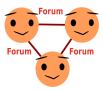
				
Real model	63259	87629	446313	3242331
Random model	495783.8	1425245.3	1423512.7	494989.6
fraction of real vs. null model	0.12	0.06	0.31	6.55

Table 4.4: Number of different triangles in the community. FFF shows triangle of Forum edges, MMM shows triangle of Messages edges, FFM shows triangle of two Forum and one Message and MMF shows triangle of two Messages and one Forum. We measure number of triangles from a null model and we show the proportion of triangles comparing real and null model.

Results do not show the same structural balance pattern in this community. This suggested that our hypothesis about friendship and enemy concept is not true in our network. We only see over representing of FFF triangles which suggests that people are tend to make a triangle discussions.

4.11 Jaccard similarity

Two people could have common friends together even if they are not friend of each other. Specially in friendship networks it is common to have similar friends. Here we use the same concept to see if members in the community have common people to contact with in terms of sending Forum posts or messages. We use an example to describe our method. Considering part of a network like Figure 4.22. Node 1 and Node 2 have three edges in common towards nodes 3,4 and 5. By definition, *Jaccard similarity* or *Jaccard index* is the proportion number of common friends between two nodes divided by number of friends in total. In this example the Jaccard similarity of two nodes 1 and 2 equals to 3/5.

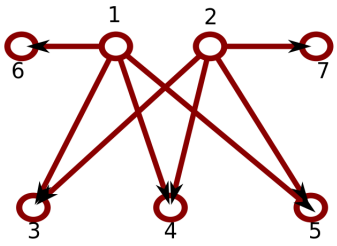


Figure 4.22: A graph sample shows nodes 1 and 2 have three edges in common with nodes 3, 4 and 5.

We measured the Jaccard similarity in Forum network and Messages network. Further we compare Jaccard similarities between Forum and Messages network. This will help to understand: 1. To what extend two members in Forum page are more tend to post comments to same people. 2. In terms of Messaging, to what extend members have same people to send messages to.

In the first step we measure Jaccard similarity in each network Forum and Messages. In general for both Forum and Messages the similarity between members is low. It means that it is not very common that two members only send posts or messages with high number of similar contacts. Figure 4.23 shows Jaccard similarity in Forum posts.

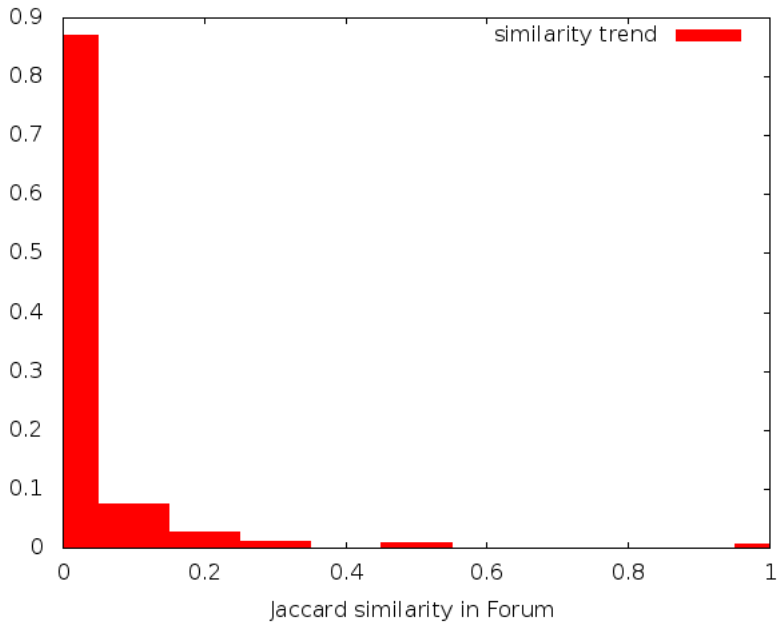


Figure 4.23: Jaccard similarity for Forum posts. The x-axis shows Jaccard index and the y-axis is the proportion of people in each bin.

Similarity in Messages, for 99% of population is between 0 to 0.1 which means similarity in Messages is highly low. We zoom in for 2% of the population to be able to observe the details. See Figure 4.24.

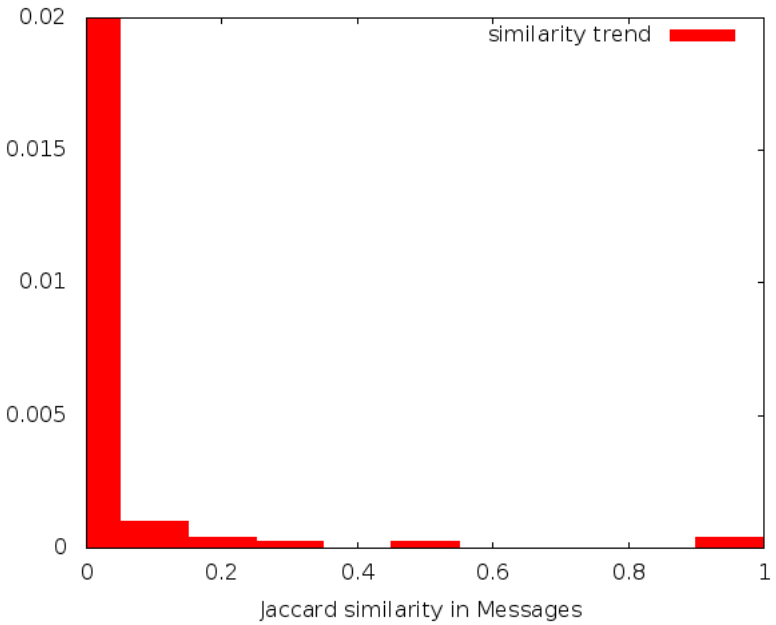


Figure 4.24: Jaccard similarity for Messages. The x-axis shows Jaccard index and the y-axis is the proportion of people in each bin. We only focus on 2% of population to be able to see the clear trend.

In the next step we compare the similarity between Forum and Messages to see if there is any difference between these two together. Figure 4.25, shows the Jaccard similarity comparing Forum and Messages. The Forum posts is in x-axis and Messages is in y-axis.

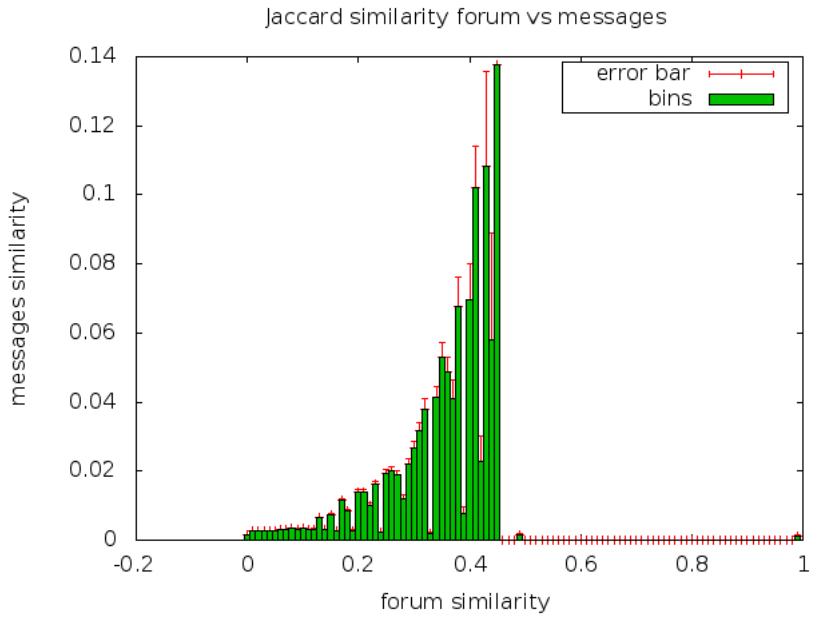


Figure 4.25: The Jaccard similarity comparing Forum and Messages. The x-axis shows Forum similarity and the y-axis shows Messages similarity.

Figure 4.26shows the Jaccard similarity comparing Messages and Forum. In this figure the x-axis is Messages and y-axis is Forum.

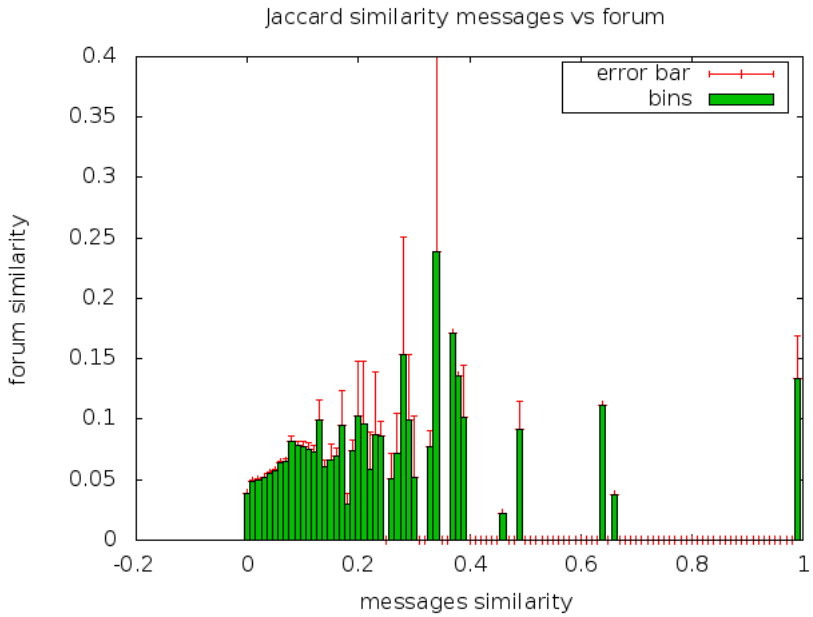


Figure 4.26: The Jaccard similarity comparing Messages (in the x-axis) and Forum (in the y-axis).

4.12 Community structure, stability and dynamic

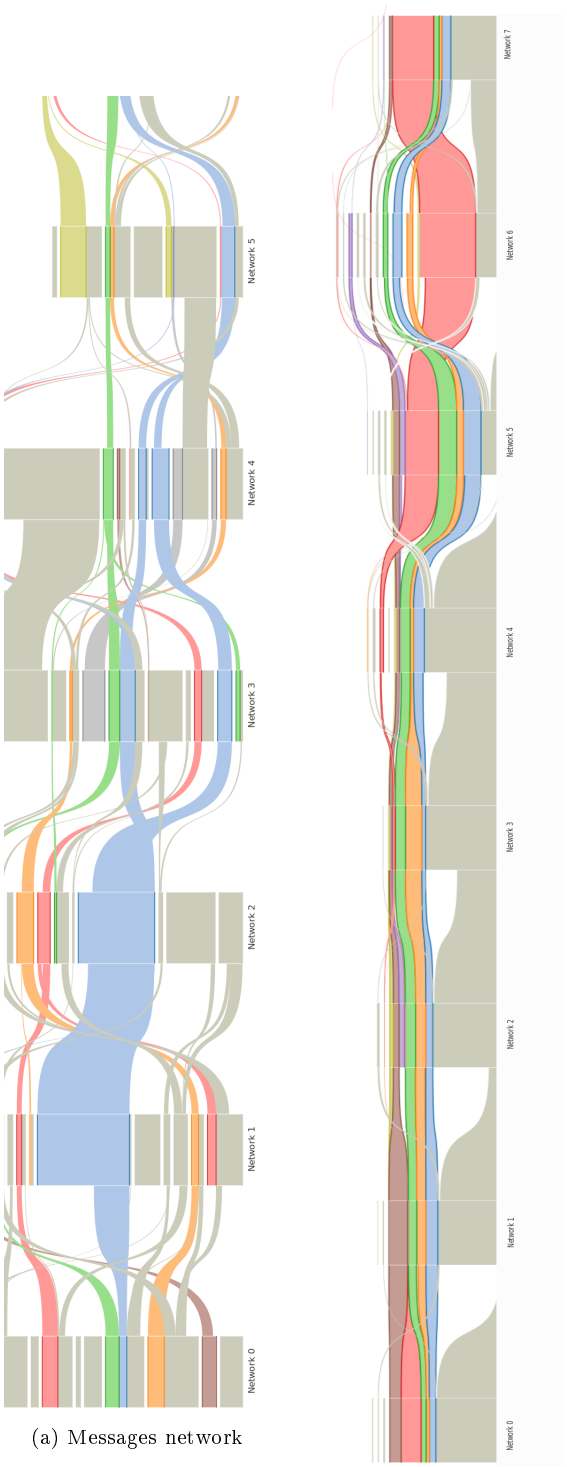
A social network consists of communities of people who make groups and interact with each other. Such examples in society are friendship groups, groups of classmates, groups of colleagues and so on. In society groups are forming, maintaining or disappearing and reforming over time. Our network as mentioned before, is a community of people who are interacting with each other due to their interests to specific field, say movie. In this case, it is somehow obvious to observe these community structures in the network.

To be able to see community structures in our network, we need to find a proper mechanism to do it for us. Here we use map generator as a tool to make clusters in the network [24]. Map generator is a random walker simulation, walks in the network depends on number of links between people. This means if there are many links or contacts between a group of people in the community, random walker spends more time in that community, therefore it generates a cluster. By doing this over the whole network, community structure will be emerged.

Our network changes over time. New members join the community and new clusters are formed or old member leave the community and old clusters will disappear. Some members meet each other in the community and depends on their movie tastes or other reasons, communicate more with each other. To

make it clear and less messy, we take N number of people who are presented in the network as time goes on. In other word, each time step, we take a snap shot of the community as long as it contains N nodes. In each snap shot there are some repeated nodes which have been before, but there are also new nodes that has just appeared in the network. Then we generate map equation and put snap shots together to see how clusters emerge or disappear or develop over time.

As an example, we took seven first snap shots of Forum and Messages networks. In each time step, we took 100 nodes or members who appears in the network. From Figure 4.27, one can see the evolution of clusters and groups in both networks, Forum and Messages. Furthermore, one can notice differences between group formation and evolution between public Forum and private Messages. In Figure (a) concerning Messages, there are more diversities in terms of clusters than in Forum network. However in Messages, Figure (a), groups in general, have shorter life time than groups in Forum, Figure (b). In Forum, there are few groups who are maintained over time which tells about groups of members who are actively participating in discussions and they prefer to discuss with specific members.



(a) Messages network

(b) Forum network

Figure 4.27: Community structure over time by using map generator. Figure (a) shows clustering evolution for messages and Figure (b) shows clustering evolution for Forum data set. Here we only took the first seven snap shots of the communities and in each snap shot, 100 nodes exist.

Chapter 5

Summary

We studied temporal network of an online movie community in Sweden. This network consists of two modes of communications depending on whether members are sending private Messages or posting comments in Forum pages to each other.

First of all we showed the emergence of community structure by looking at network properties such as clustering, assortativity, degree distributions and so on. In general the network is more likely to be disassortative. Reciprocity is high in this network. The degree distribution is following power law with cut off in Forum network.

Since our network is temporal, we measured activity and interevent time of individuals in the community and we saw daily and weekly routines of members. Response time both in Forum and Messages networks have similar pattern. Probability of responding time is higher in about 15 minuets after receiving a message. Except for the first hour that the response time in Forum is higher than Messages, both are following the same patterns. Regarding to the activity of members, we saw a correlation between activity of members in both modes which means people who are active in public Forum are also active in sending Messages. However the activity between two members of the community has anti-correlation in Forum and Messages, which reflects the fact that people have different connections when it comes to private and public communications.

The decreasing rate of interevent time is faster in Forum than Messages networks, suggesting differences in information life-time. The time interval should be short enough for participating in Forum discussions otherwise information gets old and there is no point to reply late. However in private email communications, there is more time delay in replying messages.

In terms of Jaccard similarities, members in Forum and Messages have low similarities. This shows less social force in online community to make and maintain groups. Results also show as similarity in Forum increases, similarity in Messages also increases but not vice versa.

We compare our data sets with social balance theory. We observed similarities and differences. Out of four patterns which are suggested by balance and

unbalance structures in society, three of patterns are observed in our community but one of them does not completely fit, suggesting the differences between way of interactions in society and our online community. Furthermore, we visualized our temporal network and took snap-shots to see community structures and its evolution. We observed differences in group patterns and time-evolution of groups depending on private and public communications.

Bibliography

[1] The bootleg survival guide: the privacy bubble. www.expatica.com, 2007.

[2] P.W. Anderson. More is different; broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047), 1972.

[3] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435, 2005.

[4] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 1999.

[5] Charles H. Bennett. *Complexity, Entropy, and the Physics of Information.* Addison-Wesley, 1990.

[6] Arnab Chatterjee, Sudhakar Yarlagadda, and Bikas k. Chakrabarti. *Econophysics of wealth distributions.* Springer, 2005.

[7] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdet. Scale-free topology of e-mail networks. *Physical Review E*, 66(035103), 2002.

[8] D. Gale. A theorem of flows in networks. *Pacific J. Math.*, 7:1073–1082, 1957.

[9] Nigel Goldenfeld and Leo P.Kadanoff. Simple lessons from complexity. *Science*, 284:87–89, 1999.

[10] Petter Holme. Network dynamics of ongoing social relationships. *Europhysics Letter*, 64:427–433, 2003.

[11] Petter Holme, Christofer R. Edling, and Fredrik Liljeros. Structure and time-evolution of an internet dating community. *Social Networks*, 26:155–174, 2004.

[12] Petter Holme and Jing Zhao. Exploring the assortativity-clustering space of a network's degree sequence. *Physical Review E*, 75(046111), 2007.

[13] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2002.

[14] David Layzer. *Cosmogenesis: The Growth of Order in the Universe*. Oxford University Press, 1991.

[15] Sungmin Lee, Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Title: Exploiting temporal network structures of human interaction to effectively immunize populations. e-print arxiv/1011.3928, 2010.

[16] Rosario N. Mantegna and H. Eugene Stanley. *Introduction to Econophysics*. Cambridge University Press, 1999.

[17] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.*, 27:415–44, 2001.

[18] Mark Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20), 2002.

[19] Mark Newman. Mixing patterns in networks. *Physical Review E*, 67(026126), 2003.

[20] Mark Newman. Power laws, pareto distributions and zipfs law. *Contemporary Physics*, 2006.

[21] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[22] Mark Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(035101), 2002.

[23] Jean pierre Eckmann, Elisha Moses, and Danilo Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *PNAS*, 101(40):14333–14337, 2004.

[24] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105:1118–1123, 2008.

[25] Alan Said, Ernesto W. De Luca, and Sahin Albayrak. How social relationships affect user similarities. In *Proceedings of the ACM IUI'10 Workshop on Social Recommender Systems*, 2010.

[26] David Sherrington. Physics and complexity. *Phil. Trans. Roy. Soc. A*, (368), 2009.

[27] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

[28] D. J. Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[29] Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *PNAS*, (1004008107), 2010.

[30] Stanley Wasserman and Katherine Faust. *Social network analysis*. Cambridge University Press, 1994.

[31] George Kingsley Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. George Routledge & Sons, 1936.