Proyecto Final de Carrera (PFC)

TÍTULO: Relación entre la Topología de Last.fm y el comportamiento de los usuarios en el medio social
AUTORA: Paula Sampietro Pascual
DIRECTOR: Christian Doerr
SUPERVISORES: Siyu Tang - Maarten Clements
FECHA: 09-12-2009

Network Architectures and Services Group
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

TÍTULO: Relación entre la Topología de Last.fm y el comportamiento de los usuarios en el medio social
AUTORA: Paula Sampietro Pascual
DIRECTOR: Christian Doerr
SUPERVISORES: Siyu Tang - Maarten Clements
FECHA: 09-12-2009

Resumen

Actualmente, la mayoría de las redes sociales en Internet son comunidades de personas que comparten actividades e intereses, o que están interesadas en explorar las actividades e intereses de otros usuarios.

El estudio tanto de la topología de estas redes sociales como del comportamiento de los usuarios en la red es fundamental para poder mejorar las características funcionales de la red. La red social Last.fm nos da la oportunidad de obtener diversa información pública de los usuarios: amistades, preferencias musicales y etiquetado de música.

El objetivo principal de este proyecto es determinar si existe alguna relación entre la actividad social de los usuarios (amigos) y el uso de las funciones específicas de la red social. En el siguiente documento, primero estudiamos la topología de la red en términos de distribución nodal y simetría de enlaces, y después relacionamos la distribución nodal con la actividad en la red.

A lo largo del documento extraemos una serie de interesantes conclusiones. Usuarios que escuchan música no popular tienden a tener más amigos que los usuarios que escuchan música popular. La causa de esta tendencia se debe a que los usuarios con gustos no populares están obligados a compartir sus intereses con sus amigos y buscar nuevos grupos a través de los perfiles de otros usuarios. Por el contrario, los usuarios con gustos populares no tienen estas necesidades y consecuentemente tienen menos amigos en la red.

También llegamos a la conclusión de que los usuarios que suelen etiquetar la música que escuchan (*tagging*), también suelen tener más amigos ya que están explotando más la red social en todos los sentidos y están ayudando a otros usuarios a encontrar la música con facilidad.

Con el conocimiento de cómo los usuarios actúan en la red social podremos desarrollar nuevos algoritmos que mejoren las actuales características de las redes. Un algoritmo que mejorase el inicio del proceso de hacer amistades o publicidad personalizada dependiendo de los gustos musicales son dos ejemplos de posibles mejoras que se podrían desarrollar en un futuro.

TITLE: Relation between the Topology of Last.fm and user behaviour in social media
MASTER DEGREE: Master of Science in Telecommunication Engineering (ETSETB)
AUTHOR: Paula Sampietro Pascual
DIRECTOR: Christian Doerr
SUPERVISORS: Siyu Tang - Maarten Clements
DATE: 09-12-2009

Abstract

Nowadays, most Social Network Sites (SNSs) are focused on online communities of people who share interests and activities, or who are interested in exploring the interests and activities of others.

The study of the topology of these networks and the behaviour of users are essential to improve the networking features. Last.fm gives us the opportunity to collect different data from users, such as their friends, music preferences, listening activity and tagging activity. Then, we can compare the topology of the social network made up of friendship relations and the activity of users in Last.fm.

The goal of this thesis is to determinate if there exists any relation between the social activity and how users take advantage of the site's social network functions. In the document, firstly we study the topology of the network in terms of node degree and link symmetry and secondly we relate the node degree with the network activity.

We found that users who listen to obscure artists have more relationships than users who listen to popular artists. They need to find new obscure music from other profiles and share their interests with their friends. Consequently, those users are more interested in the social activity of the network. In contrast, users who listen to popular artists have few friends because they do not need to go thought the profile pages to find music.

There is also a relation between the topology and users who tag their tracks. Most of them have large indegree and consequently more social activity. This means that they are exploiting the social network site and helping other users to find music.

With the knowledge about how users interact with the network, we can develop new algorithms to improve the networking features, i.e. a friend request start algorithm to help new users to make friends in the network or targeted advertising for users with different music tastes.

Index

Index of Figures, Tables and Equations

Chapter 1. Introduction

Social network sites are becoming more popular because users always have access to the internet. Also resources for large scale data-analysis have become affordable. Therefore, these web services can be tailored towards individual users to provide them with interesting content at any time. Then, as the popularity of social network services is constantly rising, new uses for the technology are emerging. Nowadays, most social network services are focused on online communities of people who share interests and activities, or who are interested in exploring the interests and activities of others.

However, social network sites exist with different goals. Some networks focus on the social aspects, creating and maintaining relations, (MySpace and Facebook). Other networks focus on professional relations (LinkedIn). Moreover, some networks have the primary goal to distribute home-made content (Flickr, YouTube) or to share (semi) professional content (Last.fm).

The analysis of these networks, their topology and the users' behaviour are essential to improve them and study their evolution. The knowledge of how users interact within the network might help us to develop new algorithms for recommendations, search activity and friend request start or just targeted advertising for users. These improvements are really interesting from investment and system design point of view.

This thesis is focused on the analysis of the Last.fm social network. We have chosen this network over all existing social networks because we can collect different information about users, such as friends, music tastes and tagging activity from the site operator using the Last.fm API to get it. With this information we can analyze the friendship graph and users' activities.

1.1 Problem Description

The main goal of this research is to study the relationship between user activity in social media sites and the topology of the network. Firstly we analyze the topology of the Last.fm friendship network using graph measurements like the degree distribution. Secondly we study user activity, including the preferences of users, listening activities and tagging behaviour. Finally, we study if any relation between the topology and activity of users exists and we propose some ideas to improve the networking features from the results of our analysis.

This document is organized as follows. Chapter 2 provides some background on graph theory, complex network models and power law generators. Chapter 3 describes the methodology used to collect the dataset from Last.fm website. The next chapters are focused on our measurements. Chapter 4 analyzes the topology of our network, Chapter 5 studies the crawling process and tries to simulate the social network and Chapter 6 shows the results of the relation between topology and user activity. Finally, Chapter 7 summarizes the conclusions of this work and Chapter 8 details the future work that can be done in this area.

Chapter 2. Related Work

This chapter contains background information about social networks, graph theory, the Albert-Barabasi model and existing Power Law generators.

2.1. Social Network Sites

The aim of this section is to provide a background about social networks. We begin with a brief overview of online social networks. Firstly, a definition of social network sites is given. Then we describe the characteristics of those networks. After that, we give an overview of the social networks' evolution. Next, an explanation about how Last.fm works is presented and finally we discuss some important related work done in this research area.

2.1.1. Definition and features

Social network sites (SNSs) are defined as web-based services that allow individuals to (1) construct a public or semi-public profile, (2) create a list of other users with whom they share a connection, and (3) view and go through their list of connections [1].

Profiles contain user information, such as their name, date of birth, nationality and friends' list. After joining the social network, users can search other users in the network, and make connections. Those links do not necessarily mean friendship; some of them are based on similar interests and activities. Most SNSs require bi-directional confirmation for connections, but some do not. The public display of those links is a crucial component of SNSs because users are enabled to go through the network graph by clicking through the friends' lists and consequently, create new connections. Finally, friends can share content and communicate between them leaving public or private messages.

One important research done about social networks was The Small World experiment conducted by Stanley Milgram in 1969 [2]. The Six Degrees of Separation model was included in the experiment. This experiment imagines the population as a social network and attempt to find the average path length between any two nodes. The experiment developed a procedure to count the number of ties between any two people based on sending letters with instructions around the USA and study if they arrived to the target. The researchers concluded that people in the United States are separated by about six people on average.

Thirty years later, the first recognizable social network site was created according to the definition of SNSs. It was called SixDegrees.com and was created in 1997. This network allowed users to create profiles, list their friends and go through the friends lists. Although each of these features existed in other sites, SixDegrees.com was the first to combine these features.

From 1997 onward, many new social networks have appeared. Furthermore, online communities who share interests and activities are growing. Some examples of these new kind of SNSs focused on media sharing are Flickr, MySpace, Youtube and Last.fm. In next section, characteristic features about Last.fm are presented.

2.1.2. Last.fm

Last.fm is a UK-based online social network created in 2002. It is focused on playing songs via a web radio platform and compensating artists for playing their music. 30 million people use this site and over 280,000 artists and 7 million tracks can be played [3].

Through the study of the Last.fm database we have the opportunity to analyze the friendship network topology, users' preferences in music and tagging activity. The main advantage we have with Last.fm is that we can collect data from their users with a simple application programming interface (API) provided by their web page.

Last.fm users can generate a profile page which includes basic information such as their user name, date of registration, total number of tracks played and also a section for public messages. Profile pages are visible to all, together with a list of top artists and tracks, and the 10 most recently played tracks. Profile pages can also include lists of friends, weekly top users with similar music tastes called neighbours, favourite tags, groups and events. An optional customizable playlist may also be added, with tracks that the user wishes to share or promote.
A Last.fm user can build up his musical profile using any or all of several methods: by listening to their personal music collection on a music player application, on a computer or an iPod with an Audioscrobbler plugin, or by listening to the Last.fm internet radio service. All songs played are added to a log from which personal top artist/track bar charts and musical recommendations are calculated. They call this automatic track logging *scrobbling*. Once enough songs have been scrobbled, Last.fm creates a list of musical neighbours who have similar tastes. Users can exploit these neighbours to find new music. They can also add these users to their friends list and listen to their custom radio stations.

Within Last.fm, not all users take advantage of the sites' social network functions. For some, the ability to stream music, keep track of their personal music charts, tag songs, share content and receive music recommendations are their sole motivations for using the site.
In contrast, more socially-motivated users mark others as friends and search users and their profiles through the network. Recent empirical research and theoretical development has emphasized how internet use fits into everyday communication across multiple media [5]. These studies have found that online communication does not substitute for face-to-face conversation or other forms of communication, but supplements (and perhaps even increases) offline interaction. Last.fm provides several communication platforms for those interested in using the site socially, including writing publicly-visible messages on others' profiles, sending private personal messages, and participating in site-wide discussion forums.
Signing up and creating a user profile on Last.fm is free, and so is using most of the features with the exception of the radio, which is a subscriber feature for €3.00 per

month in most countries except USA, UK and Germany from March 30th 2009 [4]. In return users will get unlimited access to Last.fm Radio, free browsing and streaming, no advertising, the ability to view recent visitors to ones' own profile page and priority on Last.fm server. Our data was collected before that need of subscription so there were no differences between user behaviour from different countries.

2.1.3. Studies in social networks

We select two papers because they are especially related to our work and now we briefly summarize their study and explain how both papers are related to our work.

Firstly, we discuss the work by Cardillo, Scellato and Latora [18]. They studied the topology of the scientific coauthorship network. In this network two scientists are considered connected if they have coauthored one or more publications together. They found that the graphs were characterized by assortative indegree–outdegree correlations, and a power-law dependence of the clustering coefficient on the node degree. Although they were not analyzing a social network site, measurements on the topology and conclusions are similar to our network.

Secondly, we refer the reader to the paper by Mislove [13] called "Measurement and analysis of online social networks". This paper will be our principal reference because it presents a large-scale measurement study and analysis of the structure of multiple online social networks. Their results confirm the power-law, small-world, and scale free properties of online social networks. For that reason our study will be much related to their paper. Moreover, they work with large-scale measurement with the majority of the dataset of these networks, so it is a good reference for us because we have only a small proportion of all the data of Last.fm. We compare our findings to their results to see whether our data is representative for a full social network. Then we extend their work by relating our network statistics to actual user behaviour in terms of tagging and music preference.

Now, we present the graph theory we need for understanding measurements and analysis done in the thesis. All the information is also collected from other literature.

2.2. Graph Theory

In computer science, *Graph theory* is the study of graphs, referring them as a collection of vertices or nodes and a collection of edges or links that connect pairs of vertices. For this study, we need some graph-theoretic data structures as a way to store information. Moreover, we need some tools to analyze the data, such as how to calculate the size, density or degree network features and some metrics to study the behaviour of that kind of networks [6]. All this theory is described in this section.

2.2.1. Adjacency Matrix

We will start describing the most common way to store data of a network, the

Adjacency Matrix **(A)**.This form of matrix in social network analysis is composed of as many rows and columns as there are users in our dataset, and where the elements represent the links between the users. The matrix is binary, meaning that if a link exists, a one is entered in a cell; if there is no link, a zero is entered. It is called an adjacency matrix because it represents who is next to, or adjacent to whom in the social space. By convention, in a directed graph, the source of the link is the row and the target of the link is the column.
An asymmetric matrix represents directed links (connections that go from a source to a receiver). That is, the element $\mathbf{A}_{i,j}$ does not necessarily equal the element $\mathbf{A}_{j,i}$. If the connections that we represent in our matrix are undirected links, the matrix is necessarily symmetric, meaning that the element $\mathbf{A}_{i,j}$ is equal to element $\mathbf{A}_{j,i}$.

Nevertheless, this way of storing data is not too suitable for our work because we are processing networks of around 300.000 users, and a matrix of 300.000 * 300.000 boxes will exceed our limits of memory. However, there is another way to store data called Adjacency List.

2.2.2. Adjacency List

The *adjacency list* contains the links between the nodes of the network. The adjacency list can be written in several ways. It contains a line for every link in a two-column form. In every row there are two numbers, representing nodes that are connected.
We take advantage of this way because our networks are not completely connected, and there are many possible links between two users that do not exist. Then, our adjacency list will have less problems with memory.

2.2.3. Metrics

Finally, in order to understand the properties of networks, topological metrics can be calculated on the dataset. In the next paragraphs these topological metrics are introduced. The list, which is based on a literature study [6], contains the most generally accepted and used metrics.

Size and density

Usually the *size of a network* is indexed simply by counting the number of nodes. In any network there are (*N * (N-1)*) unique ordered pairs of users, where *N* is the total number of users in the network. If we had undirected links the number would be *(N * (N-1))/2,* since the relationship $\mathbf{A}_{i,j}$ would be the same as $\mathbf{A}_{j,i}$. The number of logically possible relationships then grows quadratically as the number of users increases linearly. Furthermore, the proportion of all existing links compared to the number of possible links is known as the *density of the network*.

Indegree and outdegree

Differences among individuals in how connected they are can have big consequences for understanding their attributes and behaviour. More connections often mean that

individuals are exposed to be more observed. Highly connected individuals may be more influential to other users.

Users may have many or few links and they may be sources of connections, sinks (users that receive connections, but don't send them), or both.
The degree is important because it tells us how many connections a user has. For directed links, the sum of the connections from the user to others is called the *outdegree of the point*. If the user has high outdegree, he is observing many other users' profiles and preferences.

Looking at the users as sinks or receivers of information, the *indegree of the point* is the number of users that are connected to the one we are focusing on. Users that receive a lot of connections could also suffer from information overload or noise.

For undirected links, each node simply has a degree, as we cannot distinguish indegree from outdegree.

Average degree

The average node degree [7] gives information about the connectivity of a network. Networks with higher average node degree are stronger connected and they seem to be more robust. Although it is one of the most basic characteristics, it can not provide enough information by itself since totally different networks might have the same average degree. The average degree is computed by:

$$\bar{k} = \frac{\sum_{i=1}^{N} k(i)}{N} \tag{2.1}$$

where N represents the total number of nodes in the network, and $k(i)$ is the degree of node i.

Degree distribution

The nodal degree distribution [7]

$$P(k(i) = x) = \frac{|I_x|}{N} \tag{2.2}$$

is the probability that a randomly selected node has degree x. I_x is the set of nodes with degree x. It provides information about how many nodes I_x of a given degree x are in the network. Node degree distribution is the most widely used topological characteristic since it gives more information than the average degree. Knowing the degree distribution it is possible to get the average degree.

Correlation coefficient

In statistics, the correlation coefficient [7] is a measure of the strength of linear

dependence between two variables. The result obtained is equivalent to dividing the sample covariance between the two variables by the product of their sample standard deviations (Equation 2.3). That coefficient ranges from -1 to 1. In that range, a value of 1 shows that a linear equation describes the relationship between the variables perfectly and positively. A value of -1 shows that all data points lie on a single line but that one variable increases as the other decreases. Finally, a value of 0 shows that there is no linear relationship between the variables. This statistic will be useful when relating indegree and outdegree of users. The correlation coefficient is computed by:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \tag{2.3}$$

where n is the number of users, i is the ith user in the network and X and Y are the variables indegree and outdegree respectively.

2.3. Barabasi Model

In this section we introduce the concept of the Barabasi Model. A proper knowledge of this model is required to thoroughly understand the node degree distribution in social network sites. This section is focused on the model description and properties.

Power law degree distributions have been observed in different complex real world networks. *Barabasi-Albert* [8] is a model for evolving networks. The scale free networks are the ones whose degree distribution (Equation 2.4) follows a power law. The power law implies that the degree distribution of these networks has no characteristic scale. The probability distribution of the node degree in a graph with power law characteristics is represented by:

$$P(k) \approx k^{(-r)} \tag{2.4}$$

where, r is the degree exponent.

Barabasi [8] introduced the concept of growth and preferential attachment of nodes. Nodes with larger degree are more likely candidates to be attached by new nodes. The probability of being connected is proportional to the degree *k(i)* of the node. The resulting graph contains a power-law degree distribution.

The model starts with a number m_0 of fully-meshed nodes. It is important to mention that $m_0 \geq 2$. New nodes are added to the network one by one. Each node is connected to m existing nodes in the network. An existing node is chosen to be connected to the new node with a probability proportional to the degree of this node. Hence, new nodes are preferentially linked with the most connected nodes of the network. The probability p_i that a new node is linked with i is (2.5):

$$p_i = \frac{k(i)}{\sum_{j=1}^{N} k(j)} \tag{2.5}$$

where *k(i)* is the degree of the node *i.*

We are facing an inhomogeneous network where most nodes have very few connections and only a small number of nodes have many connections.

Power law degree distribution is found in many natural and artificial networks such as scientific collaboration or World Wide Web [8].

2.4. Power Law Generators

Finally, we discuss related work on power law generators, which we will use in Chapter 5. We distinguish two papers where existing power law generators are explained and compared.

In [19], a review of three existing generators that produce topologies characterized by degree power laws is presented. Power Law Random Graph (PLRG) is a generator based on curve fitting proposed in [16]. Given a number of nodes and a power law exponent, PLRG assigns degrees to nodes and then, randomly matches degrees among all nodes. We should take into account that this generator can produce self loops and duplicate links and we should eliminate. Secondly, BA power law generator is proposed in [8] as a generator that follows the Barabasi model. Starting with a small network core, at each step one of two operations is chosen, (i) adding a new node along with m links, or (ii) adding m new links without a node. This model was extended in [20] adding a new operation consisting of choosing m links and re-wiring each of them.

In [10] all those generators are presented and compared, and some more are explained. Havel-Hakimi [21] is a deterministic generator included in the curve fitting family. Given a degree sequence, the resultant graph is always the same. The algorithm ensures that high degree nodes will always be connected with high degree nodes. Takao algorithm [22] is also a generator of the curve fitting family, which requires a degree sequence as input and is also deterministic.

In chapter 5 we will discuss the characteristics that the simulated network will need and then, we will determinate which generator will reach better our goal.

Chapter 3. Dataset

In this chapter we describe firstly the dataset and the methodology that we use to collect it and secondly the steps followed for the data processing.

3.1 Description of the dataset

One of the main reasons that we choose Last.fm in our study is that we can collect different information about users, such as friends, music tastes and tagging activity from the site operator using the Last.fm API.

We define two crawling processes, firstly the graph for the social relations, which includes friends of users and secondly, the graph for the tastes of users, which includes the top 50 artists, the top 50 tracks, the top 50 albums, the tagging activity and also friends of users. We will call these graphs respectively as the *social graph* and the *taste graph*.

We started by doing a breadth first crawl of the Last.fm network, collecting information about friend relationships. These friend relationships are connections in the network between two users who want to share information and their profiles.

Information of 304,122 users was crawled for the social graph before stopping the process, including their friend relationships. Our data was written in text files, with the name of <username>.txt. Each line inside the .txt file contained a friend's name. That data was crawled during a period of 28 days, starting in May 28th 2008 and finishing in June 24th.

Next, we started our second crawl to collect Friends, Tracks, Albums, Artists and Tags of each user. We maintained the same user order as the one used to crawl the social graph, starting from the same first user but stopping it when we had data of 119.130 users. These users form a subset of the larger graph because we started the crawl from the same user.

For Tracks, Artists and Albums, we have a list of the top 50, including a playcount with the number of instances that the user listened that track, artist or album. In contrast, in Tag files we have a list of tags most used by users, with a limit of a maximum of 1000 different tags per user, but only 46 users of the whole dataset have that maximum. Those files also have a playcount with the number of instances that users tag with each word. An example of track information is presented in the following, although this structure is followed also by artist, album and tag information:

```
<track>
      <artist mbid="df765d93-621c-437f-99fe">Cornelius</artist>
      <name>Music</name>
      <mbid></mbid>
      <playcount>70</playcount>
      <rank>1</rank>
      <url>http://www.last.fm/music/Cornelius/_/Music</url>
</track>
```

We observe in this XML file the name of the track, the artist who plays it, the playcount meaning the count of instances that the user listened to that track, the rank of that track in the users' profile and the URL address where we can find it.

In fact, the files created contained only the top used elements, so we do not have the entire user profiles. We used XML files to store the data. All these data presented is from a crawl conducted on June 9th 2008, and it finished on June 24th.

We have the social graph, with friend information about 300.000 users, and the taste graph, which includes a subset of the first one and it is formed by music preferences and tagging activity about 120.000 users.

3.2. Processing the dataset

To obtain an adjacency list, we first convert each user name (in string) to a unique identifier (ID). After the conversion, the name of username.txt is converted to userid.txt. Furthermore, the friends inside userid.txt are represented also by IDs. The test.awk program was created to solve that problem. Following is an example of the file UserID.txt.

UserID	UserID_2 (Friend 1)
UserID	UserID_3 (Friend 2)
…	

By concatenating all files after conversion, an adjacency list is obtained for the social graph. This new file called AdjacencyList.txt contained 2.334.267 different links between users.

We should bear in mind that in our data there exist more links, but the dataset which we are working is a part from the whole dataset of Last.fm (around 30 million users), so there exist friends of our users which are not included in that network. Hence, it was decided to eliminate all those links that weren't on our dataset from files.

We should also mention that the crawling process we used in both graphs and the way we randomly stop the process causes an anomaly in the topology. Many users have indegree while zero outdegree. However, this fact is explained in detail in section 5.2.

Chapter 4. Topology of the Last.fm Friend Network

This chapter focuses on the analysis of the topological property of the Last.fm friend network. The goal is to analyze the behaviour of the Last.fm network by evaluating the degree distribution. We present our results and findings in the following sections.

Recall that all our steps would be carried out for two networks, i.e. the taste graph and the social graph which contains the taste one. Since the taste graph is a subset of the social graph, it is interesting to find out whether these two graphs have the same property. This problem is commonly known as the sampling problem, which allows us to understand a large and complex network (e.g. the social network) by proper sampling a sub-graph of the network.

So, we have the social graph, with around 300,000 users, and the taste graph, which includes a subset of the first one and it is formed by 120,000 users. All steps would be detailed first for the social network and then for the taste one. Finally, we will compare both results and validate the taste graph for the user activity analysis.

4.1. The node degree distribution of the social graph

First of all, in order to obtain the measurement results, data from 304,122 Last.fm users was collected. After the brief overview of the dataset, we notice that the most popular user has 3,481 friends (UserID 228,719). In contrast there are 64,752 users without any friend, which implies 21.29% of empty files.

The adjacency list was created as explained in chapter 3. The graph contains 2,334,267 different links between users. If the maximum number of links that could be present between 304,122 users equals 92,489,886,760, the fraction between the total and the maximum number of links, known as *link density*, equals 0.0025%. This value is appropriate because the link density of a power law graph should be very low, since there are many low-degree nodes.

In our study, the node degree is chosen as a metric to evaluate the topological property of the Last.fm network, so that we can see whether the Barabasi model can be applied testing the existence of a power law in the degree distribution. We also compared music preferences with the number of friends for each user and then study if relations between tastes and social activity in the network exist. This study is presented in chapter 6.

Hence, for the degree distribution analysis, two AWK programs were made. Each program generates an output file called Indegree.txt and Outdegree.txt respectively. Both files contained two columns, the first one with the UserID and the second one with the degree. The number of head endpoints adjacent to a node is called the *indegree* of the node and the number of tail endpoints is its *outdegree*.

For further analysis we chose MATLAB, because we could use existing functions that are integrated in MATLAB.

In the following, we describe the results that we have obtained. The mean of the social graph indegree equals 7.6754, with a standard deviation equals 23.098. The number of users with indegree zero, which are referred to as the *source* since it is the origin of each of its links, equals 49,604 (16.3%). The most popular user has 2,832 incoming edges. The probability distribution of indegree is depicted in the Figure 1. We can observe that the majority of the users, 85.6 %, and thus the most dominant indegree size, is smaller than 14. To fit the frequency presented in Figure 1 with a line curve, we use the methodology provided in Newman [9]. The fitting function returns the scaling parameter α and the starting point of the fitting $x_{min.}$
A power law exists and it begins at the 14th degree, with an exponent α equals -2.59. That means that the graph is following the scale free model.

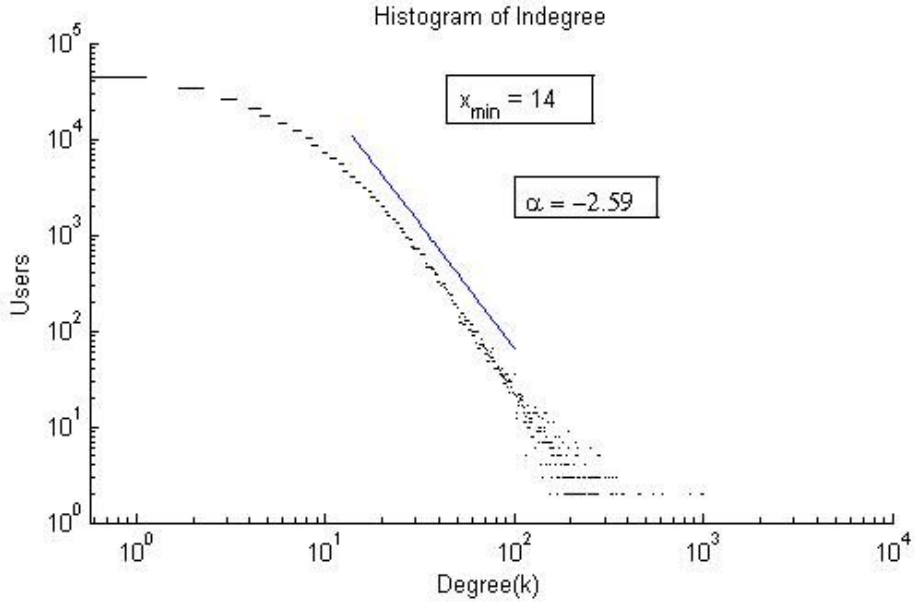

Figure 1. Indegree histogram with logarithmic axes.

Next, we present our findings for the outdegree. The mean of the outdegree equals 7.6754, the same value as the indegree, and it has a standard deviation equals 25.3134. The number of users with outdegree zero, called *sinks* because the user is the target of its entire links, equals 78,922 (25.9%). The most popular node of the Last.fm social network has 3,104 friends. Shown in Figure 2, the distribution of the outdegree has similar behaviour as the indegree. The majority of the users, 85.3 %, and thus the most dominant outdegree size, is smaller than 16. Using the same java program, we found that the tail of the histogram obeys the power law with α equals -2.59 and x_{min} equals 16.

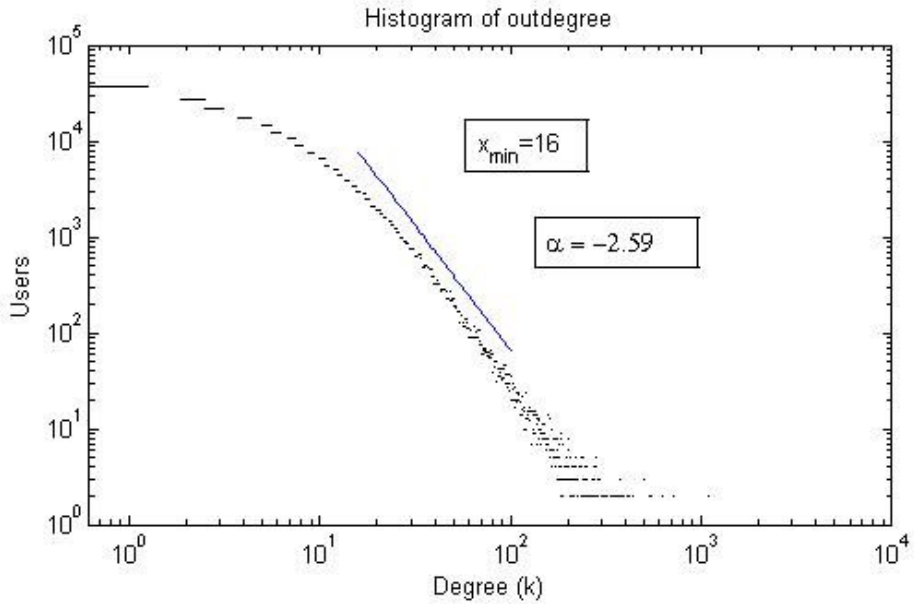

Figure 2. Outdegree histogram with logarithmic axes.

From the above analysis, we have found similar results for the indegree and outdegree in the Last.fm network.

In the following, we study whether there is a correlation between the indegree and the outdegree in the Last.fm network. First of all, we measure the differences (in absolute value) between the indegree and outdegree for each user.
The mean value of the difference between in and outdegree equals 1.8402 and the standard deviation equals 7.8586. There are 173,267 users who have the same indegree and outdegree, which means 57% users of the total social graph.

Next, we analyze the difference between the indegree and the outdegree without the absolute value. We can see in Figure 3 that an 11.6% of the nodes have more indegree than outdegree and 31.4% of the nodes have more outdegree. Theoretically, the total indegree should equal to the total outdegree. However, in Figure 3 both graph sides seem to be really different. The reason is that users with more outdegree are a larger number of users, but in contrast the ones with more indegree are more spread in the axis.
Moreover, the majority of the users, 91.9%, have the difference between ±5. That is, the in and outdegree tend to be really similar for the whole network.

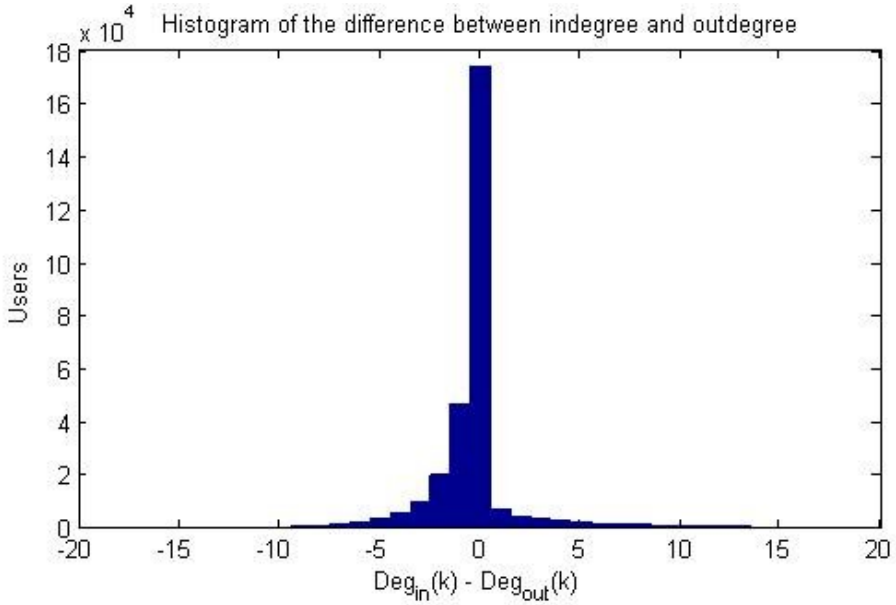

Figure 3. Histogram of the difference between indegree and outdegree with linear axes.

In Figure 3 we observed the existence of many users with the same indegree and outdegree. To find out whether the abnormal behaviour found in Figure 3 is caused by *inactive users*, it was thought interesting to eliminate the users without activity, which are the ones with indegree and outdegree equal zero. There were 15.6% of users in that situation. As we can see in the Figure 4 the graph without inactive users is not so different from the one with them. We also notice that the distribution is not symmetric. As we explained before, if we make the difference between degrees in total terms, it equals zero. Hence, we found that there are more users with more outdegree than the ones with more indegree. However, users with more indegree are more spread and differences are bigger.

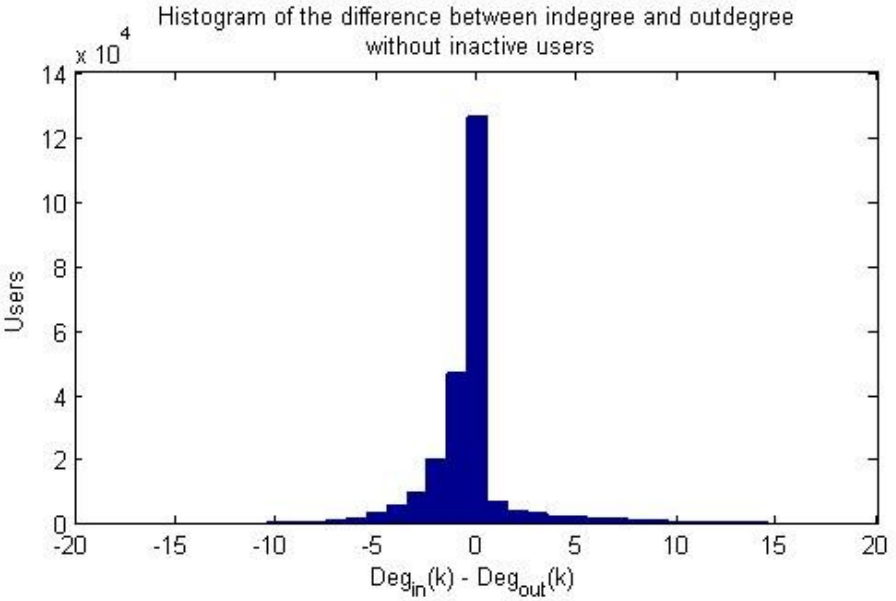

Figure 4. Histogram of the difference between indegree and outdegree with linear axes and without inactive users.

4.2. The node degree distribution of the taste graph

In this section we study the friendship network with only the users that we have information about their music preferences. As we mentioned before this new network is a subset of the social graph analyzed above. The network is formed by 119,130 users and 989,521 edges between users, which means a link density equals 0.0069%. The indegree and outdegree are measured and the results are explained below.

Firstly, we found that the mean of the taste graph indegree equals 8.3062, with a standard deviation equals 22.0666. There exist only 91 users with zero indegree, which is less than 0.1% of the whole data. The most popular user has 2.722 input edges. The distribution of the indegree is depicted in Figure 5. It is noticed that the majority of users, 88.2%, have indegree smaller than 15. The tail of the histogram obeys the power law with α equals -2.59 and x_{min} equals 15.

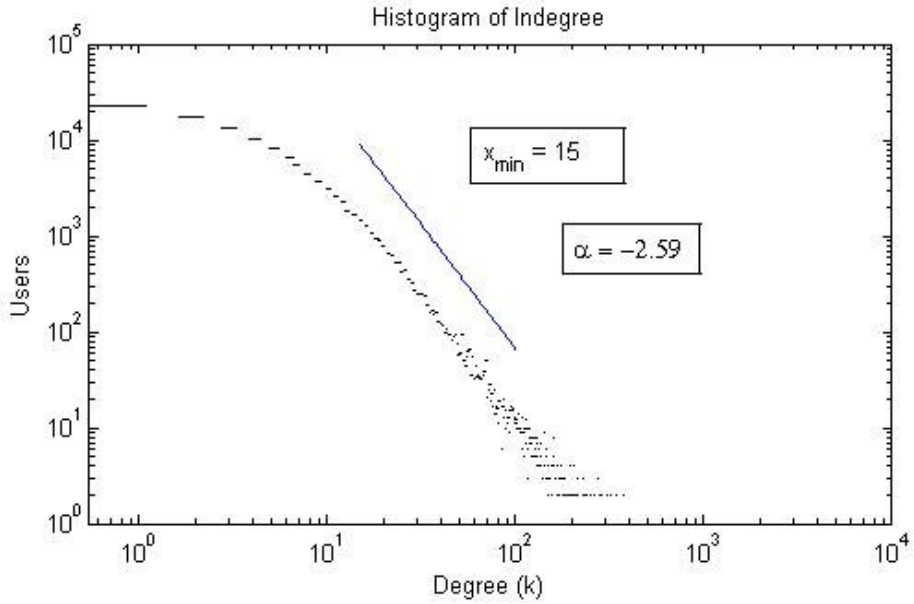

Figure 5. Indegree Histogram with logarithmic axes.

Secondly, the outdegree is also studied. The mean of the outdegree equals 8.3062, which is the same to the indegree histogram. The standard deviation of the outdegree equals 26.7531. The number of users with outdegree zero equals 30.669, which represents a 25.7% of the total nodes. The most popular user has 2,718 friends. The histogram of the outdegree is presented in Figure 6. As we can see, the majority of users, 85.7%, have outdegree smaller than 14. As explained before, this is the position where the power law begins. The scaling parameter α equals 2.42.

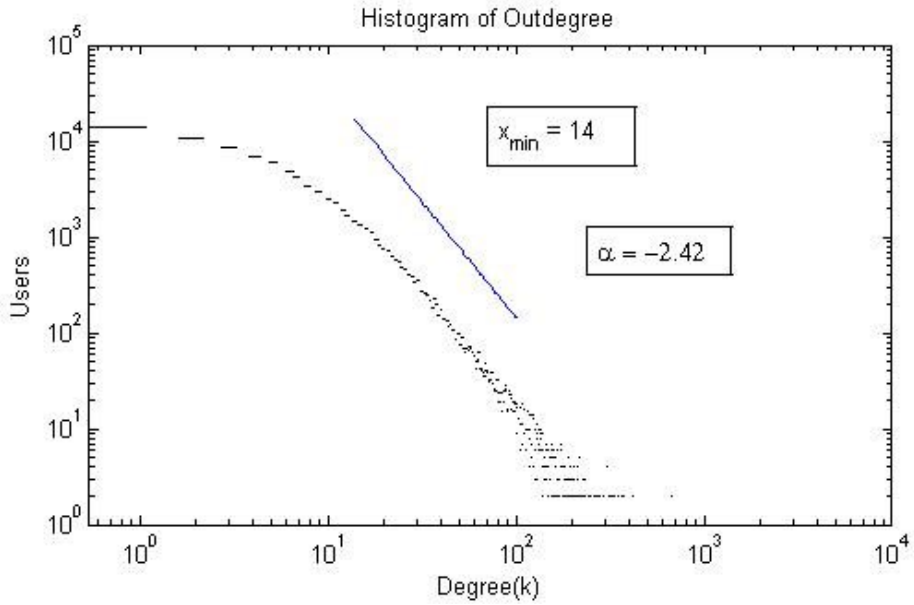

Figure 6. Outdegree Histogram with logarithmic axes.

Finally, to perform a similar measurement to investigate the differences of each user between the indegree and the outdegree, a new graph is created. The mean of the difference equals 0, as the total indegree should equal to the total outdegree. If we look below at the distribution of the difference we can see a histogram more symmetric than the one for the social graph. See Figure 7.

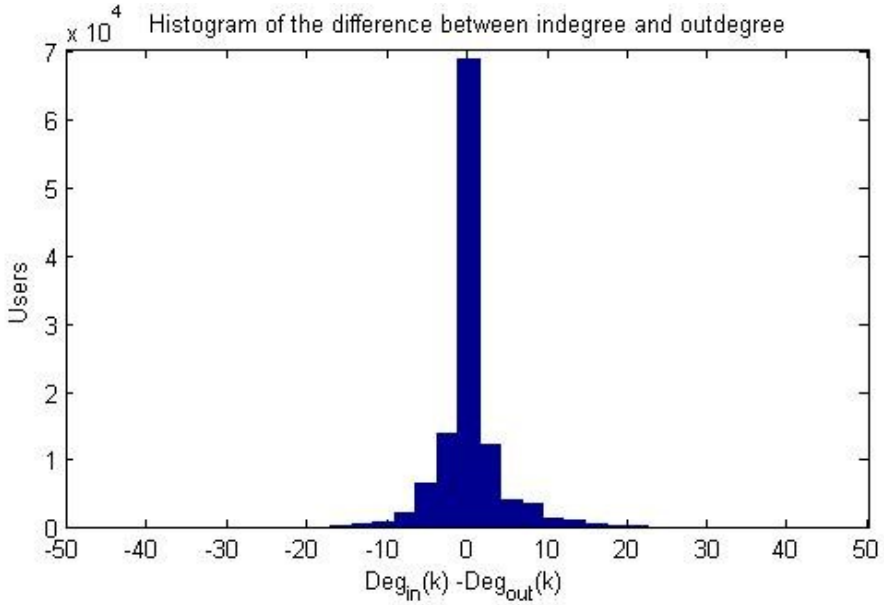

Figure 7. Histogram of the difference between indegree and outdegree with linear axes.

Same results as in the social graph were observed when extracting inactive users, although the graph is not included due to the absence of new information with it.

The comparison between two graphs is detailed in section 4.5, where the taste graph

will be validated as sufficient in terms of node degree.

In next section, another strategy to analyze the data is implemented and with it we try to solve some doubts about the skew of that graphs.

4.3. Difference between indegree and outdegree

In this section, we study the difference between the indegree and the outdegree of the Last.fm network in more detail. In Figure 8, we plot for the social graph the curves on a three-dimensional coordinate, in which the x-axis represents the indegree k, the y-axis denotes to the difference between the in and outdegree, and the z-axis is the count of instances. In Figure 9, we present the results for the taste graph, which shows similar behaviour compared with the social graph.

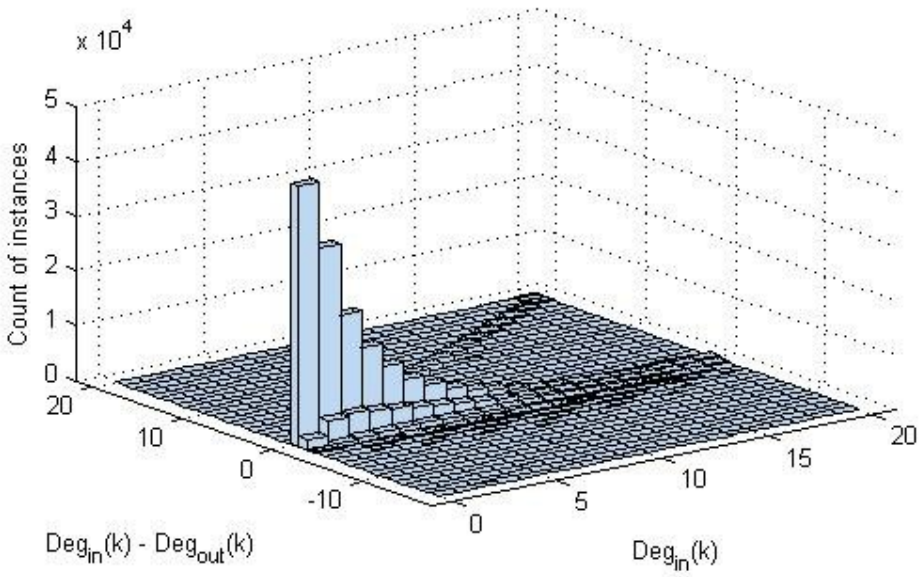

Figure 8. Histogram of the difference between indegree and outdegree vs. indegree for the social graph.

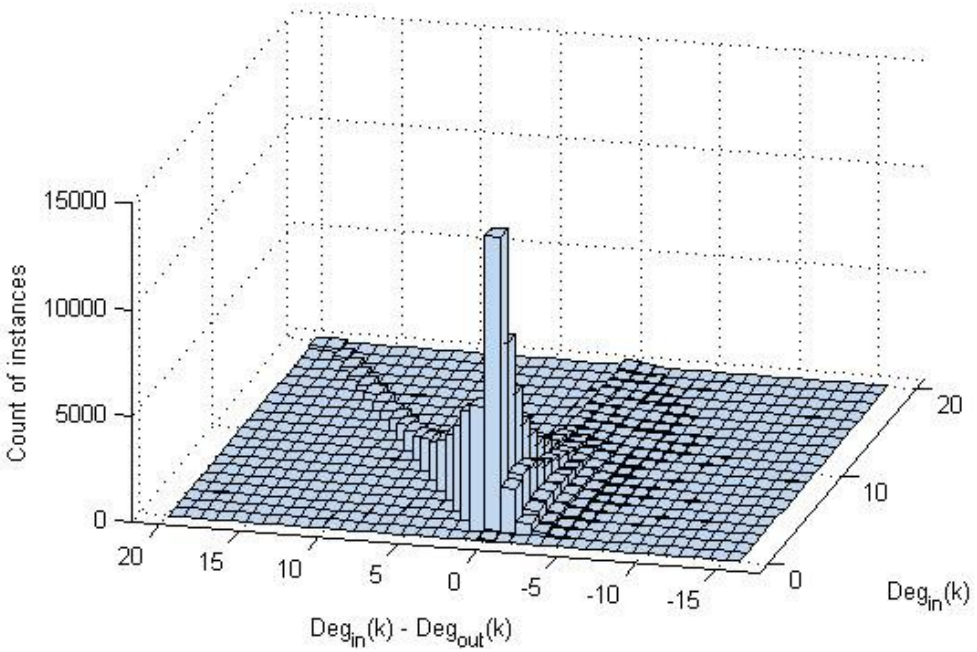

Figure 9. Histogram of the difference between indegree and outdegree vs. indegree for the taste graph.

From the above two figures, we observe similar phenomenon as found before. Most of users have same the indegree than outdegree. However, additional information can be obtained from the 3D plots.

First of all, users with outdegree larger than indegree have a fast decay. This observation is due to the fact that new users need to start friend request. One possible option to improve this situation in social networks would be to create a small model based on notifications for old and high connected users with the information of the new users. Then, if any old user accepts the connection, the new user would have an easier start friend request.

Secondly, for users with more indegree than outdegree, we find that most of them follow an unexpected behaviour. All those users have their outdegree equals zero and their indegree is larger than 0, creating a distribution in the diagonal. That is, most of users with indegree (Deg_{in}) equals x have the difference ($Deg_{in} – Deg_{out}$) equals x. In Chapter 5, we explain this anomaly with extended analysis, and we show that it is in fact caused from the crawling process.

In next section, the symmetry of our network and the correlation between the indegree and outdegree is analyzed.

4.4. Symmetry and Correlation

After studying the degree distribution, we analyzed another metrics of the network, e.g. the link symmetry and the correlation between the indegree and the outdegree of each link.

To calculate the link symmetry, we have to check the adjacency list for each link if

the symmetric link exists. For the social graph 87.5% of links of our network are symmetrical. For the taste graph, that metric equals 82.2% of links. With these values, we can conclude that in most of connections users are mutual friends.

Next, the correlation between the indegree and the outdegree for each user is evaluated by (equation 2.3) [7].

In the first case, with the social graph, we got a value equals 0.9485 and in the second case, with the taste graph, the value of the correlation coefficient equals 0.9317. In both cases, that values show the strong and positive linear dependence between the indegree and the outdegree.
It is also useful to plot the relation between the indegree and the outdegree because then we can see if they are mostly correlated and if they follow a linear regression. In the first case, it follows the equation (4.1) and in the second case, it follows the equation (4.2). Next, those relations and the linear regression fits are plotted in Figures 10 and 11 respectively.

Outdegree = 1.0395* Indegree - 0.3029; (4.1)
Outdegree= 1.1296* Indegree - 1.0765; (4.2)

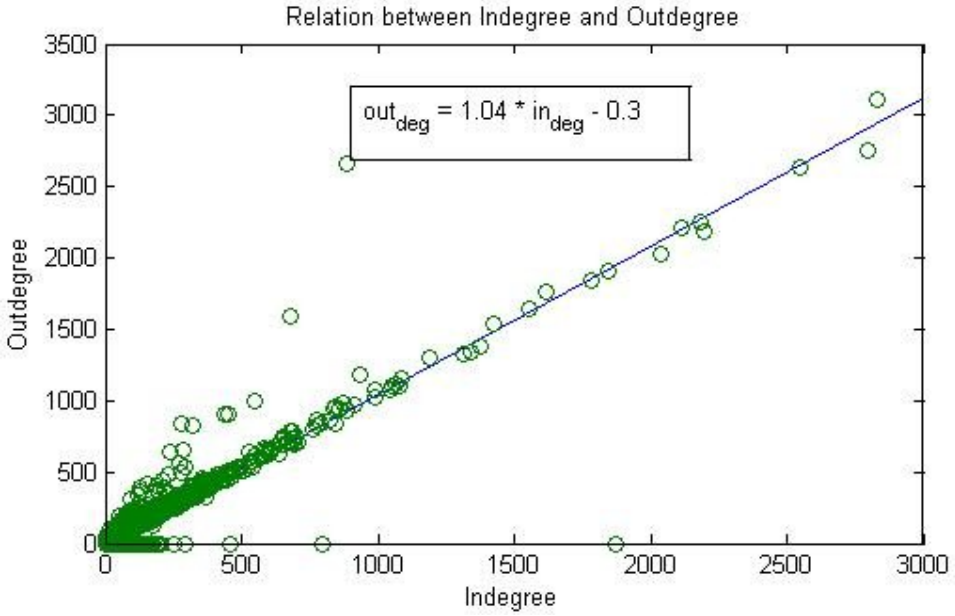

Figure 10. Relation between indegree and outdegree and linear regression fit for the social graph.

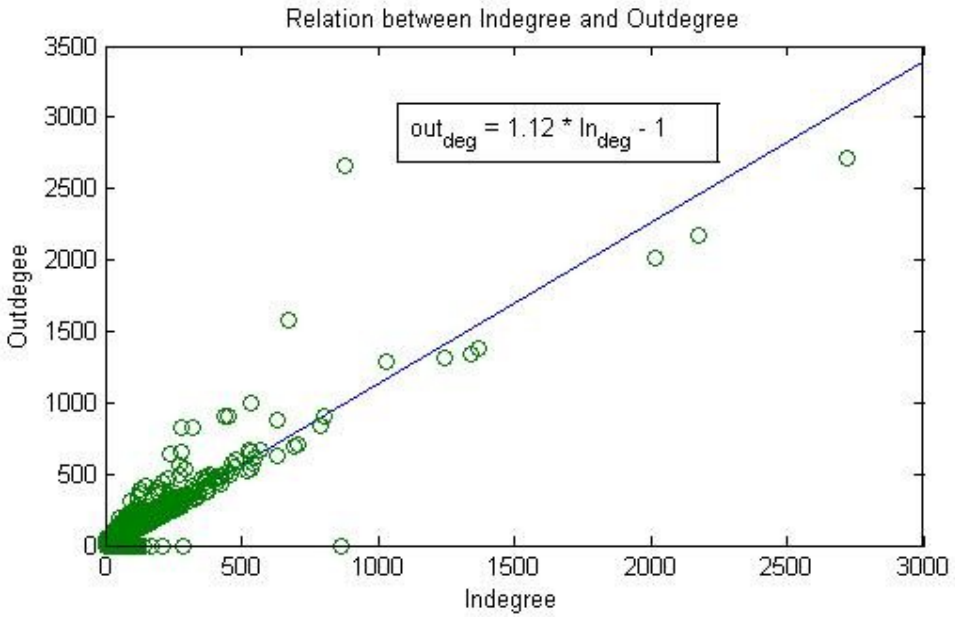

Figure 11. Relation between indegree and outdegree and linear regression fit for the taste graph.

As we can see from Figures 10 and 11, there exist a high correlation between indegree and outdegree. It means that most of users have the same indegree and outdegree, or at least similar values. Moreover, this correlation increases when we just analyze users who follow the power law. When considering the tail distribution that obeys a power law (for both the in and outdegree), the positive correlation is even stronger, as shown in Figures 12 and 13, respectively.

In the social graph, we calculate the correlation coefficient when $x_{min} \geq 14$ for the indegree and $x_{min} \geq 16$ for the outdegree, and it equals 0.9753. In the taste graph we calculate the correlation coefficient when $x_{min} \geq 15$ for the indegree and $x_{min} \geq 14$ for the outdegree. We obtain that the coefficient equals 0.9445. Both cases improve their correlation between indegree and outdegree.

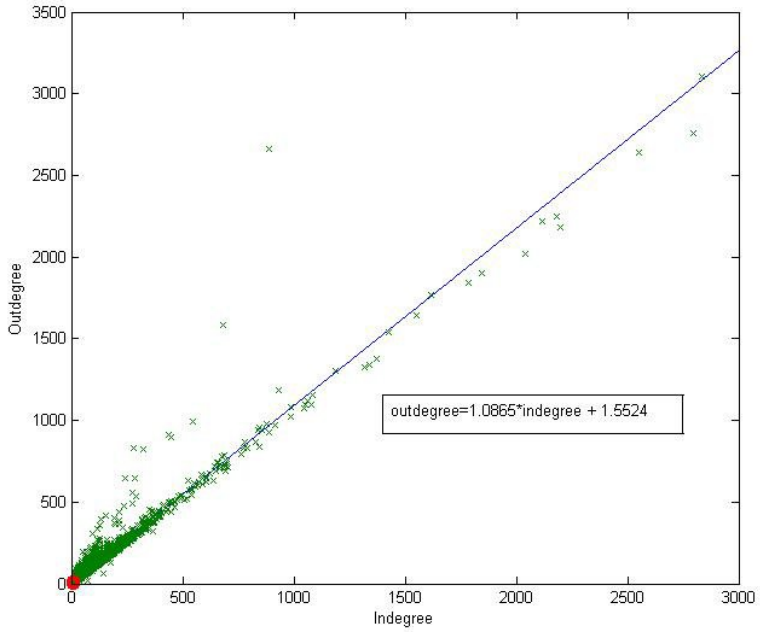

Figure 12. Relation between indegree and outdegree and linear regression fit for users from the social graph that follow the power law degree distribution.

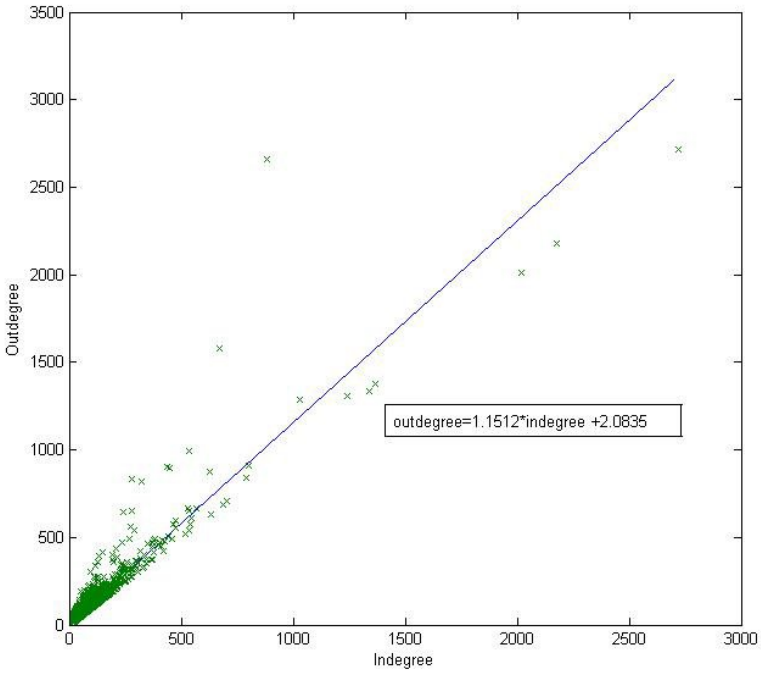

Figure 13. Relation between indegree and outdegree and linear regression fit for users from the taste graph that follow the power law degree distribution.

To sum up we conclude that there exist a bilateral behaviour between the indegree and the outdegree. However, with it we can not explain the anomaly in the topology that we found in the last section.

4.5. Validation

In the last section we will explain and validate the taste graph as a subset of the network, which will be sufficient when comparing the node degree with users' tastes in chapter 6.

As we can see in the above sections, the social graph with around 300.000 users and the taste graph which is contained in the social graph, show similar characteristic in terms of node degree and link symmetry.
Firstly, the mean values of indegree and outdegree were really similar if we compare the graphs. In both networks the mean value of indegree was the same to the outdegree. For both graphs, also power law exponents of the degree distributions were quite similar, with identically values for indegree and closed values for the outdegree. Secondly, referring to the difference between indegree and outdegree, we took same conclusions for the social and the taste graph, with the majority of users with same indegree and outdegree and users with larger outdegree than indegree having a fast decay of the distribution due to users' age. With the correlation coefficient we corroborated that a strong linear dependence between indegree and outdegree exists for both graphs. Finally, with the symmetry analysis, we conclude that most of users have friendship relations approved by both sides.

Therefore, we claim that when studying the node degree and the link symmetry, it is sufficient to analyze the subset of the larger network, i.e. the taste graph. The scale of the graph does not affect the topological property of the network.
Consequently, we can analyze only the taste graph in chapter 6 without problems of memory when comparing topological properties with user tastes.

Also mentioning that we assume that our topological results can also be extended to the full Last.fm network. The reason is based on the same idea as in last paragraph. We assume that our dataset is containing a subset of the full network, and then, our results are representatives. However, we know that the crawling process brings biases to the dataset. We visualize this idea in Figure 14, with circles that represent the different networks.

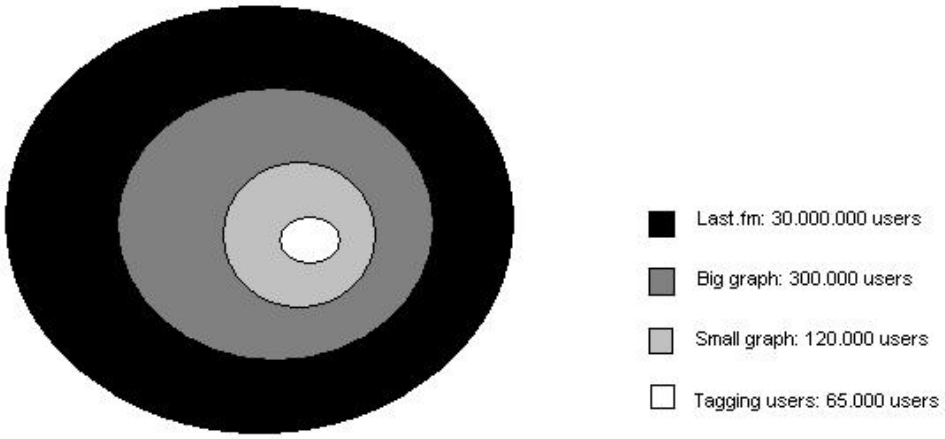

Figure 14. Circles that represent the networks used in this study.

Other possible work for the future will be to study which would be the break point of enough users for our analysis, searching the moment when results would not be representative. It would be done crawling samples from the same user with the first breath method and stopping the process in different times.

Chapter 5.Crawling & Power Law Generator

The aim of this chapter is to provide a good explanation of our anomaly in the topology found in section 4.3. We decided to follow a strategy. It consisted on creating a new network, around 10 times bigger than the social graph, which will follow a power law with some expected exponents and when it would be created, we will collect data from a number of users several times, each time beginning from different users. Then, if everything works correctly, we will be able to explain the anomaly in our crawled network.

The crawling process is based on the *breadth first method*. It consists on starting in one user, collect data from him and from his friends, then collect data from the friends of his friends, after that the friends of their friends, …, and repeating this process until we decide to stop the crawl. Consequently, the size of our network depends on two factors, the starting point of crawl and the time we decide to stop the process.

First of all, we do some literature on existing power law generators and then we program our power law generator.

5.1. Literature

Starting with the literature, we searched in internet and other sources some information about existing power law generators.
The basic condition we needed in the generator was that it could crate a graph with directed links and different exponents for indegree and outdegree. We found several generators, such as one based on the Boost Graph Library [12] and other based on the igraph library [11]. Both were programmed following the small world model. In [10] a qualitative comparison of power law generators is presented and with it we could understand how these programs work.

To explain the anomaly founded in chapter 4, we thought which parameters should be necessary for our new network. In paper Measurement and Analysis of Online Social Network [13] a large-scale measurement study was presented based on Flickr, YouTube, Livejournal and Orkut. We use for our study the power law coefficient estimates from Flickr because this network works similar to Last.fm and these values were measured with the database of the majority of users and they are representative. Figure 15 shows the values we are going to use.

	Outdegree		Indegree	
Network	α	D	α	D
Web [12]	2.67	-	2.09	-
Flickr	1.74	0.0575	1.78	0.0278
LiveJournal	1.59	0.0783	1.65	0.1037
Orkut	1.50	0.6319	1.50	0.6203
YouTube	1.63	0.1314	1.99	0.0094

Figure 15. Power Law Coefficient estimates from [13].

Using the matrix analysis program R, we just had to install the igraph library [11], a free software package for creating and manipulating undirected and directed graphs, chose the parameters for the number of users and exponent and save that network in the optimal format, which in our case will be an edgelist.

We used two different algorithms, first the barabasi.game( ), a free-scale graph generator according to the Barabasi-Albert model and then the aging.prefatt.game( ), an evolving random graph generator with preferential attachment and aging.

First function is a simple stochastic algorithm to generate a graph. It is a discrete time step model and in each time step a single vertex is added. We start with a single vertex and no edges in the first time step. Then we add one vertex in each time step and the new vertex initiates some edges to old vertices. The probability that an old vertex is chosen is given by

$$P[i] \sim k[i]^{alpha} + a \qquad (5.1)$$

where *k[i]* is the indegree of vertex *i* in the current time step and *alpha* and *a* are parameters given by the arguments.

In the other hand, aging.prefatt.game ( ) creates a random graph by simulating its evolution. Each time a new vertex is added, it creates a number of links to old vertices. The probability that an old vertex is cited depends on its indegree (preferential attachment), age and is proportional to

$$P[i] \sim (c\, k[i]^{alpha} + a)\, (d\, l[i]^{beta} + a) \qquad (5.2)$$

Here *k[i]* is the indegree of vertex *i* in the current time step and *l[i]* is the age of vertex *i*. The age is simply defined as the number of time steps passed since the vertex is added. *c*, *alpha*, *a*, *d*, *beta* and *b* are parameters and they can be set by the arguments.

Following all the steps we just got the expected indegree but outdegree was always 1 for both algorithms. We conclude that the Barabasi-Albert model and the package igraph [11] were not suitable for our problem because we were only able to fix the exponent of the indegree power law distribution, but in any case the exponent of the outdegree with directed links. Consequently, we decided to try with the second option.

That second option uses the Boost graph library [12], a C++ library with many graph algorithms. We create a new network with the pre-programmed function based on PLRG. This generator works perfectly for undirected links, following the input exponent, but for directed links, only the outdegree distribution is considered. Then, the indegree is disregarded and can be anything.
Finally, we realized that we could not get two different exponents for indegree and outdegree with these two options.

After analyzing the other existing generators [10,16,17,19], we discuss why the other power law generators do not fit our problem. Both Takao and Havel-Hakimi generators required a degree sequence as input and BA creates a power law associated

only to the indegree. Then, we noticed that any power law generator carry out two basic conditions to create a network similar to the Last.fm friend network: a generator with directed links and different exponents for the indegree and the outdegree power law exponents (1), and specific correlation between the indegree and the outdegree for each user (2).

Consequently, the next step was to create and program a new power law generator based on the last two conditions.

5.2. Power Law Generator

After taking the decision of programming our extended version of the power law generator, we thought that the best option will be to write a new program based on the one we had been working with. In this section we will explain all our steps and troubles.

This basic PLRG generator works easily. It starts building a vector with all the stubs of each node. Each stub is identified by its vertex index. A map is built which maintains the available nodes and if available nodes equal less than two, the iteration ends. Remember this program is the same we used in our literature based on the Boost Graph Library [12] and whose indegree is disregarded.

So now, we should modify this program considering the two basic conditions we need. The first modification is the addition of two vectors containing stubs, one for the indegree distribution and the other for the outdegree distribution. Next change in the program is related with the correlation between the indegree and outdegree of users. Consequently, the outdegree will be reordered following the linear equation (4.1 and 4.2) we found in Chapter 4 for the correlation. Then, the indegree and outdegree for a user will be correlated. These stubs will be randomly linked, connecting one from the indegree vector with another from the outdegree. Finally, the iteration ends when one of the stub vector sizes less than two.

This extended version will create networks with directed links, different exponents for indegree and outdegree and correlation between indegree and outdegree for each user.
We found that repeated links were created. It was solve adding a new condition in the loop of creating links between stubs. We also found with the new program that if we want to carry out the condition of different exponent for indegree and outdegree, consequently, different number of stubs will be created as well as an unbalanced network. Then, we always have stubs left and, depending on the exponent, they will be from the indegree or the outdegree. Moreover, if the indegree exponent is bigger than the outdegree exponent and if we analyze the created network, the outdegree distribution will work correctly but the indegree will have anomalies. The opposite case for outdegree exponent bigger works identically. It means that many users with indegree larger than 0 will have zero outdegree.

If we plot the relation between indegree and outdegree we find that this behaviour has similar appearance to our anomaly in the network topology.
In the next two graphs this behaviour is compared and we will try two explain this coincidence. The relation between indegree and outdegree for the Last.fm network

appears in Figure 16 and the same relation for the created network appears in Figure 17.

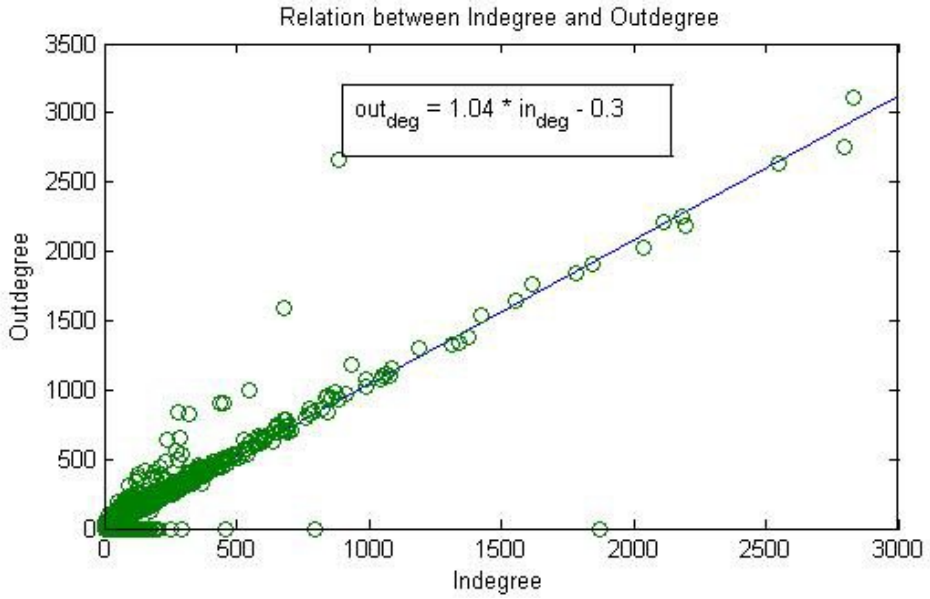

Figure 16. Relation between indegree and outdegree for the Last.fm network.

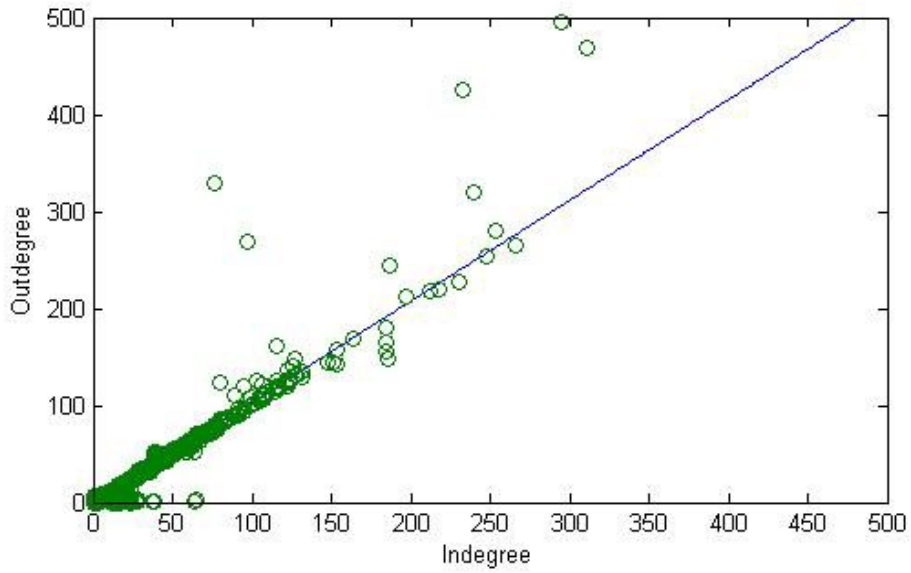

Figure 17. Relation between indegree and outdegree for the created network.

We can easily notice this similar behaviour in both graphs. The existence of a big proportion of users with indegree and zero outdegree in both cases gives us an explanation of the anomaly in our network.

If we think how our crawling process worked and we follow the crawling steps, we realize that the final step is not as clear as we thought in the first moment. Our

crawling process is based on the first breath method, so starting from a random user, we collect the information of his friends, and the friends of his friends respectively. But the problem is that the process was stopped in a random moment, deciding that the dataset was sufficient. In that moment there where lots of users that have not finished their collecting process. That is the reason why there exist too many users with indegree and zero outdegree.

We can now explain this strange behaviour. Although in a first analysis we thought our anomaly could be due to the topology of our network and the user behaviour adding friends, now we realize we were wrong and it was only due to the crawling process and this trouble produced by the stopping moment of the crawl.

We now have to know that the moment of stopping the crawling process with the first breath method will have consequences and if we want to collect the whole dataset of these users, we should include a condition to complete the collection or eliminate users from the last step.

Chapter 6. User activity vs. Network topology

In this chapter we compare measurements of the topology of the taste network with the music consumption behaviour of the users. Three different relations will be studied: the relation between the indegree and the listening activity of users (1), the relation between the indegree and the taste of the users (2), and finally the relation between indegree and tagging behaviour of users (3). This will show the differences between people who have lots of friends and others who are more interested in music, we will show if users with comparable taste also have similar indegree and whether friends tend to prefer similar tags.

6.1 Indegree vs. Play count

The next study is based on the comparison between the music listening activity and the social activity for each user. We study this relation to see whether people mainly use Last.fm as social platform or just as replacement for a traditional radio station. We compare the indegree of the social network with the play count for each user. For each user our data set contains a list of the top 50 most played tracks, and a play count for each track. We presume that the sum of the play counts in the top 50 of a user gives a reasonable estimate of the listening activity of that user. We acknowledge that this assumption ignores differences between users with a broad or narrow interest. The full listening profiles are however not available through the Last.fm API.
In Figure 18 the relation between these two variables is depicted. From this figure we can observe that there is a relation between these variables. Users with maximum indegree have little play count, and also the opposite case, users with maximum play count have little indegree.

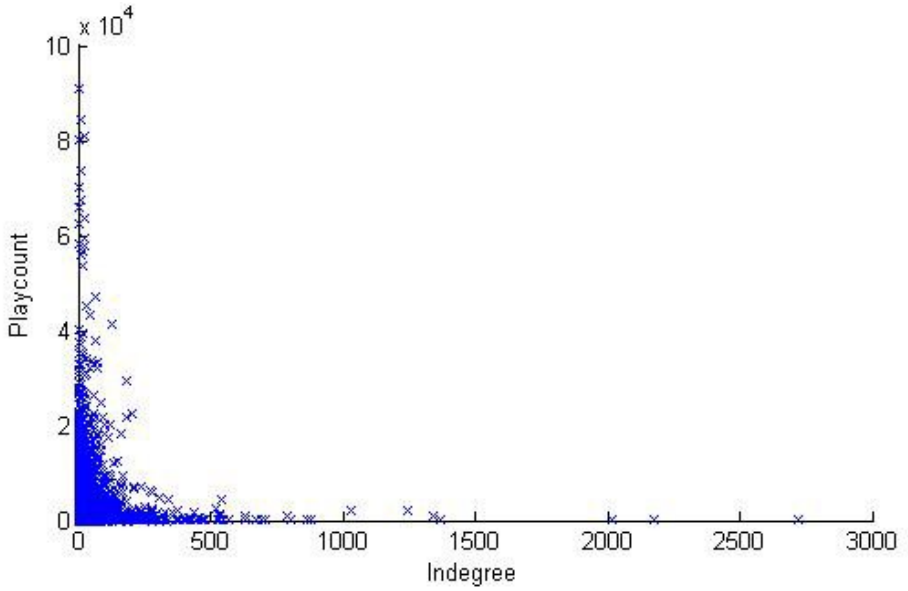

Figure 18. Relation between indegree and playcount for each user. The tail of the playcount axis is removed for readability.

To study the extreme users, the outliers of the data in both dimensions are shown in Table 1.

Playcount	Indegree
188	2178
29	2017
39	2722
623,658	4
377,741	7
304,384	1

Table 1. Relation between indegree and playcount for extreme users.

This table contains the extreme cases. First, we present the three users with the maximum indegree and secondly the three users with most playcount. We can observe a clear relation for these extreme cases. Users with most social activity have really little listening activity. And also the opposite case is observed, with the maximum playcount and relatively very few friends.
We can conclude extreme users only focus on a single aspect of the system. They either use the website as a social platform to become friends with as many people as possible, or they don't exploit the social network in Last.fm and merely use the website as personalized radio station.
We also can conclude that although there are extreme users, the majority of them are compensated, with a normal behaviour for the listening and the friendship network.

6.2. Taste vs. Indegree

Now, we will compare the indegree with the music taste of users. We follow two strategies to evaluate the assumption that if a user listens to popular music, more people want to be his friend. In the first strategy, we select the top 20 artists from the Last.fm website. Then, we find users who listen to these artists and see if they have different indegree on average. In the second strategy, we compute the average popularity of each user's top 10 artists and plot this against their mean indegree to validate our assumption.

First, we compare users who listen to same artist and then see if between them any relation exists. From the Last.fm website, we collect the top 20 artists for the week ending on June 8th 2008 [14]. We choose this week because it ends just before our crawl.

For each artist we select the users that have this artist in their top-10 most played. We then compute the mean indegree and the standard deviation of these groups of users. All these data are presented in Table 2.

Top Artists	Listeners	Indegree	Standard deviation
Radiohead	103,054	8.55	17.05
Coldplay	95,331	6.71	15.24
The Beatles	88,929	8.41	23.09
Red Hot Chilli Peppers	74,226	6.16	14.29
Muse	67,598	7.87	12.61
Metallica	67,357	6.94	18.87
Linkin Park	58,449	6.70	12.67
Death Cab for Cutie	58,280	6.89	9.95
Nirvana	57,753	6.86	14.23
Foo Fighters	55,870	6.58	14.38
The Killers	55,082	8.31	24.41
Pink Floyd	54,425	8.11	21.69
Weezer	50,917	7.03	16.83
System of a down	49,281	6.23	13.03
Led Zeppelin	49,258	7.39	19.49
Green Day	47,494	6.61	17.22
Daft Punk	47,345	8.45	16.95
Nine Inch Nails	46,482	8.72	27.11
Queen	46,209	7.19	14.30
Artic Monkeys	46,015	8.54	19.77

Table 2. Top artists from Last.fm on June 8th 2008 [14], the mean and the standard deviation of indegree for listeners of those artists.

After computing the mean indegree for people who have listened to the top 20 artists, we compare these mean values to the overall mean indegree of the network (from Chapter 4) which equals 8.30. This comparison is presented in Figure 19.

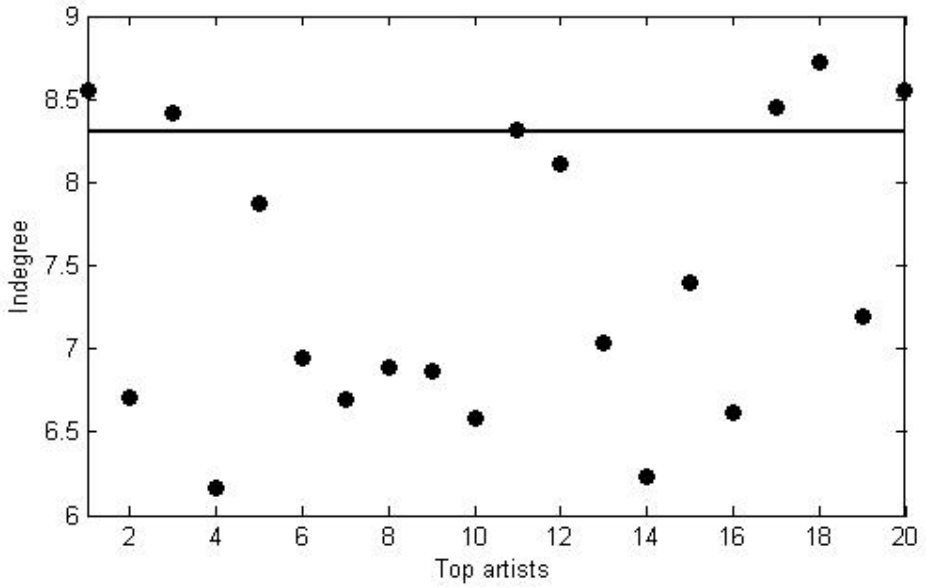

Figure 19. Mean indegree for each group of listeners of top 20 artists and the mean indegree of the whole network (equals 8.3062).

In Figure 19 and in Table 2 we can see that for 14 out of 20 user groups related to popular artists the indegree is lower than the mean indegree of the whole network. Moreover, the 6 groups of users who have larger mean indegree than the general value are really close to that value, with a difference of 0.42 in the worse case. So we conclude that the average indegree of groups of listeners is lower than the overall mean for many popular artists and then, our assumption is false. Users who listen to popular artists do not have more friends. This opposite hypothesis can be explained by the fact that users who listen to popular artists do not need to share their interests with other users and they are not interested in finding new or obscure artists from other profiles. Consequently, they are less interested in the social activity of the network and have fewer friends.

Now we follow the other strategy. It consists on compute the average popularity of the top10 artists for each user and create a scatter plot of indegree versus the average popularity. We decided that the first twenty artists in the top list from the Last.fm website are the ones that can be considered as popular, and depending on the position in the list [14], each artist will have different weight for the popularity computation of each user. Also, the artist will have different weight depending on the position of the top list of each user, meaning that the most played artists for a user will have bigger weight than the 10^{th} artist most played. This computation is explained with the next equation.

$$populatiry = \sum_{i=1}^{10} p(a \mid u) * p(a) \qquad (6.1)$$

In this equation, i is each top-10 user artist, a is the artist, u is the user, $p(a|u)$ is the weight of the position in the top track ranking for this user, meaning a value of 10 if the track is the most played and 1 if the track is the 10^{th} most played. Also $p(a)$ is the weight of the position in the top track ranking from Last.fm website. That is, a weight of 20 for the most played and 1 for the 20^{th} most played. Finally, all the popularity computations are normalized.

With this computation that ranges between 0 and 1, significant results are obtained. In Figure 20 we present the mean indegree of groups of users who have similar popularity computation. We divided the popularity of user's top 10 artists in smaller ranges and compute the mean indegree for each range.

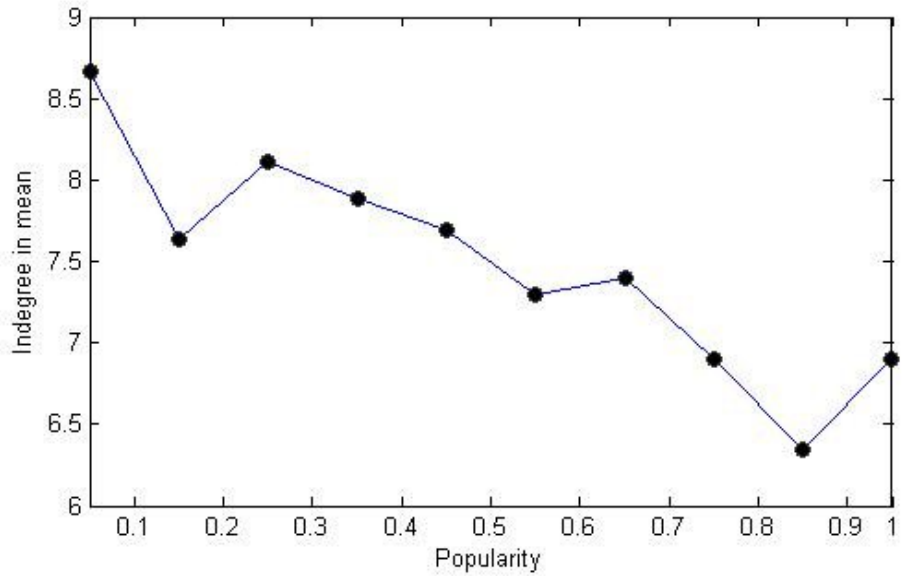

Figure 20. Mean indegree for each range of popularity computation of users' top 10 artists.

We clearly observe that users who have a larger popularity score and therefore, users who listen to more popular artists tend to have smaller indegree. Both in the graph and in the table we notice that the mean indegree decreases with popularity of artists listened by the user. This behaviour is due to the fact that people who listen to popular music can easily find this in the network and therefore do not need to exploit the social network to find and discuss the music they listen to.

Moreover, studying the relation between indegree and popularity computation of extreme users we also observe clear behaviour. Extreme cases of this relation are presented in Table 3. Users with the largest indegree do not listen to popular artists, but in contrast, users who listen to the most popular artists tent to have small indegree. This reaffirms our previous conclusions.

Indegree	Users' popularity of top 10 artists [0-1]
4	0.96
7	0.97
1	1
2178	0
1017	0
2722	0

Table 3. Extreme cases for the relation between indegree and popularity of top 10 artists of users.

Finally, we summarize this section explaining the results obtained. Both strategies confirm the theory that users who listen to popular music have less activity in the social network than the overall mean of the whole network. Users who listen to popular music can simply turn on the radio and listen to what is played. People who are interested in more obscure artists actively have to search for their music, and our results indicate that they exploit the social network to this end.

6.3. Tagging behaviour

We will now analyze the tagging behaviour of users, including general statistics in the activity of users, a comparison between the taste graph and the tagging graph and finally the differences between friends and no friends in tagging.

6.3.1. General tagging statistics in Last.fm

In this section we will introduce what is the tagging activity, study how many tags people use and finally analyze the most popular tags.

From August 2005, Last.fm supports user tagging of tracks, albums and artists to create a site-wide social classification. Users can browse via tags, but the most important benefit is what Last.fm calls *tag radio*, permitting users to play music that has been tagged a certain way. This tagging can be by genre (e.g. "rock"), artist characteristic (e.g. "female vocalist"), or any other form of user defined classification (e.g. "seen live"). Last.fm allows users to browse the most popular tags in a visualization called *tag cloud*, where the most prominent tags are typeset in a larger font (Figure 21).

Figure 21. Example of Tag Cloud from Last.fm.

In our dataset only 54.65% of the users actively tag their tracks, albums or artists. Consequently, we have a network with 65,095 active users with at least 1 tag. We analyze the number of tags per user without taking into account inactive users. We

can get a maximum of 1000 tags per user through the api, but only 46 users had them. The mean equals 18.7 tags and it has a standard deviation equals 55.28. We also can see in Figure 22 that 90% of active users have less than 50 different tags.

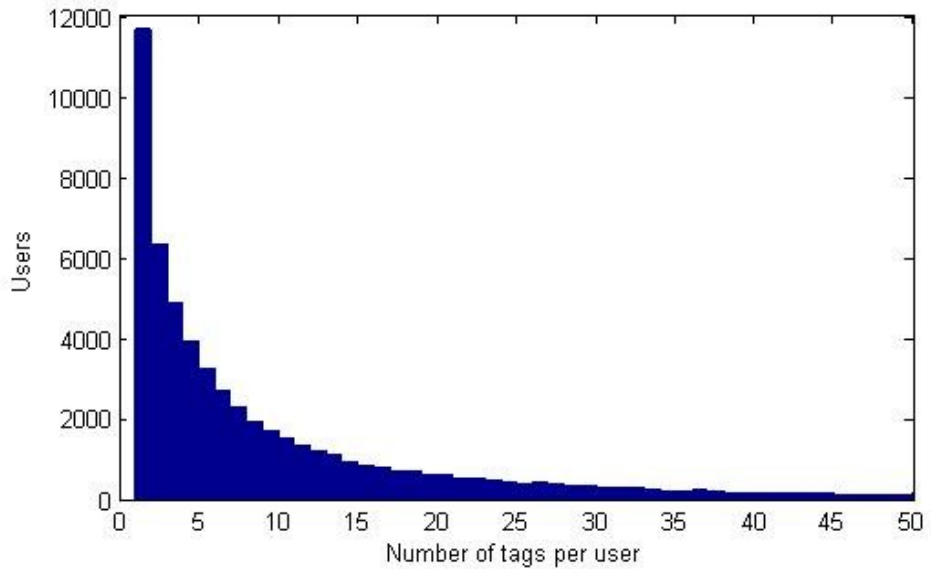

Figure 22. Histogram of tags per user without inactive users. The tail of the data is removed for readability.

To compare user preferences we are only interested in the most commonly used tags. We thought interesting to follow a strategy, based on analyzing the top 10 tags of each user. With this strategy we tried to rule out some information of the users that have many tags and then just take into account the most popular tags for each user.

Keeping on our strategy, we have a maximum of 10 tags per user and the distribution of the number of tags per each user is depicted in Figure 23.

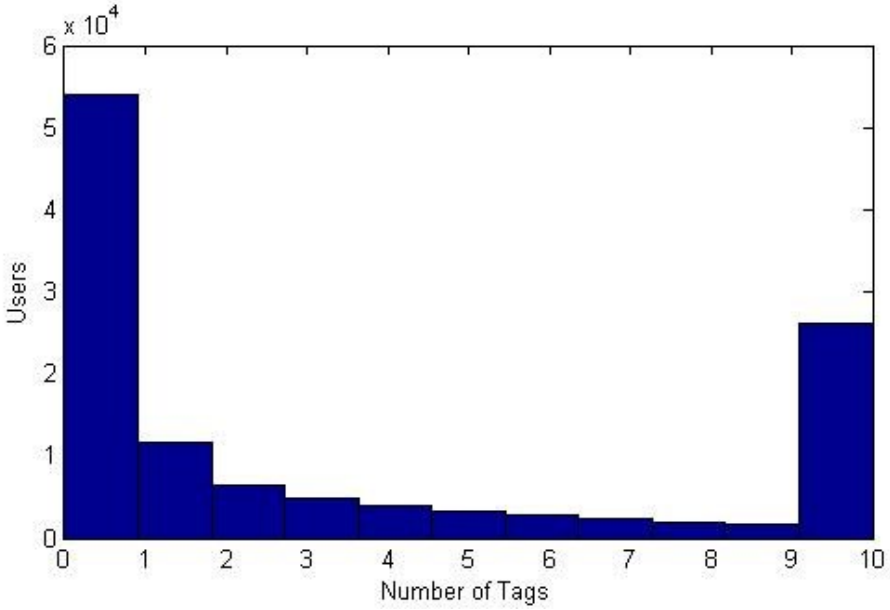

Figure 23. Histogram of the number of tags per user.

As we can see, most of users do not tag and there is an accumulation in the 10 tag bar of all the users that have more than 10 tags. Nevertheless, the mean of tags per user equals 3.3365, and after eliminating the inactive users equals 6.1062.

Next, we made a study of the most popular tags between our users and we realized they correspond to the most popular tags that Last.fm represent in the tag cloud shown in Figure 21. Table 4 contains the 10 most popular tags, with the count of instances.

Tag	Users
Rock	11,182
Indie	8.964
Alternative	7,522
Electronic	7,065
Seen Live	6,849
Pop	4,484
Indie Rock	3,993
Ambient	3,623
Female vocalist	3,400
Metal	3,393

Table 4. 10 most popular tags in Last.fm taste graph.

6.3.2. Compare tagging graph with taste graph

In this section we will create a graph based on tagging activity of users, called the *tagging graph*. We will discuss the statistics of that graph, compare it with the taste graph from chapter 4 and finally study the relation between the tagging activity and the social activity.

The first step to create the tagging graph was to relate each different tag with an identifier and then easily work with the data. Next, we created the adjacency list with all the links between users. In this graph a link means at least one tag in common between two users. Two users can have between 1 and 10 tags shared. We now create the tagging graph by counting the number of shared tags between each pair of users and then use the number of shared tags between them as the weight of the edge.

We detail some important items related to the tagging graph features. Our network is composed of 397.483 tags and 239.336.533 links between them. If the graph can have a maximum of (k*(k-1)/2) possible links, k being the number of the network users, the link density of our network equals 11.3% if we consider just the possible links between active users and 3.37% if we consider the whole network.

Next, we will analyse the distribution of link weights. As mentioned before, users can have between 1 and 10 links shared, so the weight of each link will range between 1 and 10. It is just a method to difference links which have different number of tags shared. We created Table 5 with the count of instances for each link weight and the percentage related.

Weight	Number of links	Percentage %
1	1,82E+08	75.94
2	41229520	17.22
3	11953136	4.99
4	3341625	1.39
5	841269	0.35
6	174355	0.07
7	27194	0.01
8	2844	<0.01
9	191	<0.01
10	12	<0.01

Table 5. Count of instances for each link weight and percentage related.

In short, we see that most of links have weight 1, (75.94%), and there only exist 12 user pairs that have an identical top-10 of tags.

Then, we did the study of the degree distribution of the tagging graph. Without considering the inactive users, the average degree equals 7353.5. There exist 4,918 users with degree zero (7.5% of active users), and the maximum degree of a user equals 29,933. The degree distribution is depicted in Figure 24. It has linear axes because it was not following any power law and then the logarithmic axes did not provide us any information.

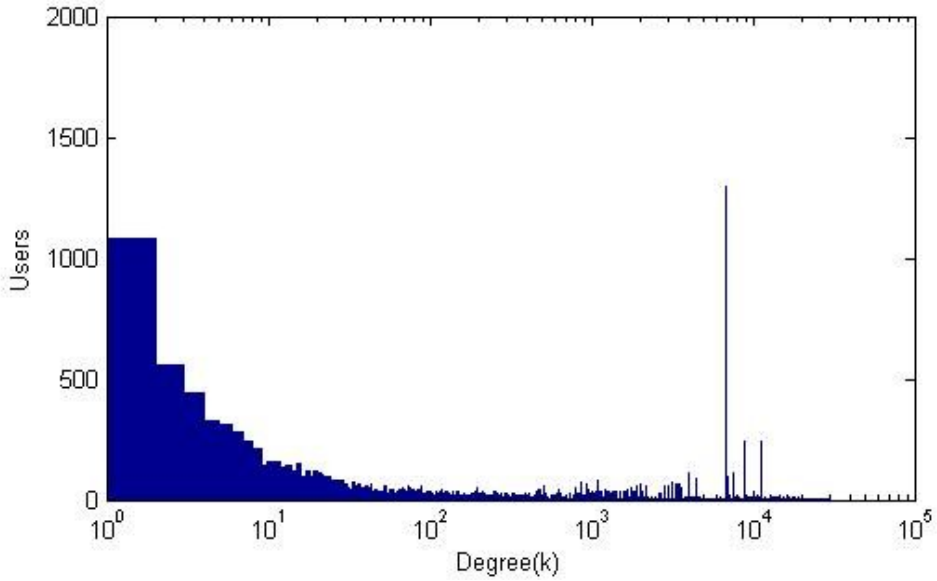

Figure 24. Histogram of the tagging graph degree.

We can see that most of the users have degree smaller than 15000, 80%. We can also see some important peaks in the graph. This is due to common tags between users that create too many links between them. These tags must be not too popular, so then most of the users that have these tags are only sharing these ones with the network. For instance, the peak we can see in the degree equal 6848 it is due to 1294 users that only

have in common the tag “seen live”, which is used by 6849 users. The next peak, in the degree 8963 is due to 241 users that only shared the tag “indie”, used by 8964 users. And another example is the peak for the degree 11181, which is formed by 246 users that only share the tag “rock”. That tag is used by 11182 users in total.

We compare the taste graph and the tagging graph and we see some strong differences. Firstly, the tagging graph contains undirected links because if two users share one tag, a link is created between them without source and target. In contrast, the taste graph contains directed links. Secondly, the degree distribution of the tagging graph does not follow any power law contrary to the taste graph, because the tagging graph is not based on the small world model. Also the average values for node degree are completely different. For the taste graph the mean indegree equals 8.30 and for the new graph the mean degree equals 7353.5. And finally, the link density of both graphs is extremely different. Since the tagging graph has a link density equals 3.37%, the taste graph has a link density equals 0.0069% for the same number of nodes. So we conclude that these two graphs are completely different and similarities do not exist in any way.

Finally, we compare the tagging activity with the indegree based on number of friends. Our assumption is that users who tag their tracks and use many different words in tagging are exploiting more the social network site. The tagging activity helps other users to search music in the network. So if a user tags many songs with different words, he will contribute to the site and probably will have more social activity.
We decided to calculate the mean indegree for users with closed number of tags. We divided the dataset and plotted the mean indegree value for each group. From 0 to 100 tags per user, the data is divided in smaller groups because most of users have less than 100 different tags. All these results are presented in Figure 25.

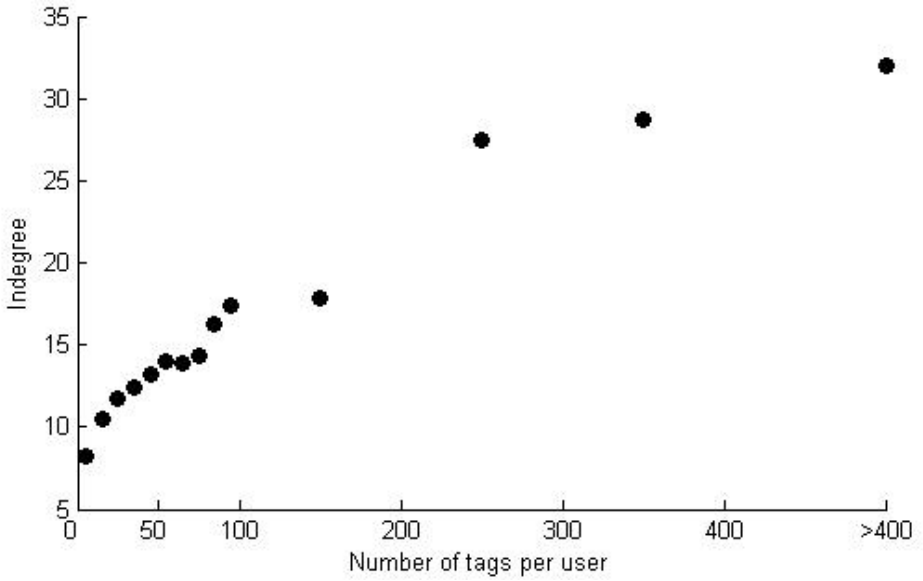

Figure 25. Mean indegree for each range of number of tags per user.

Although inactive users are not presented in Figure 25, we calculate their mean indegree and it equals 6.15. Then, users without tagging activity have less mean indegree than active users. Moreover, the graph clearly shows a tendency. We observe that users with larger number of tags tend to have larger indegree. Consequently, we can conclude that our assumption is true. Users with more tagging activity have larger indegree and then, also more social activity in the network. This means that they are exploiting more the social network site and helping other users to find music. Probably, this behaviour is related to people with obscure music interests because as we concluded in section 6.2, users with rare preferences also have larger indegree. Moreover, users who listen to popular music do not need to help others to find that music.

6.3.3. Compare friends tag graph to non-friend tag graph

Last but not least, we create two separate graphs, one for friends and one for non-friends, and we compare them to see if there exists any difference in the tagging behaviour.
We analyzed if there existed more common tags between friends than between users without any relationship in the taste graph. We observe in Figure 26 that friends have more tags in common and the weight is larger in those links.

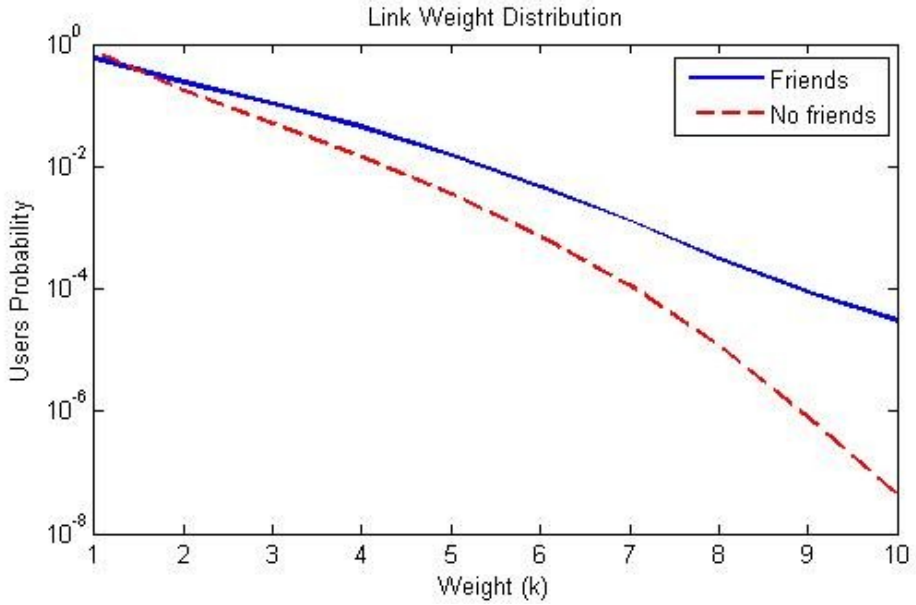

Figure 26. Comparison of link weight distribution for friends and no friends.

We conclude that a link in the tagging graph is more likely to be between a pair of users with a link in the taste graph. The reason of this behaviour is that friends often have a common vocabulary or music taste.
This conclusion is interesting because we can use it to improve the social network site creating new algorithms. If friends have similar interest they could be exploited to predict recommendations for a user. Then, if a user is searching for new music and we show him what his friends are listening, we are probably recommending him useful information.

Chapter 7. Conclusions

In this chapter we are going to show the main conclusions obtained after the development of the project.

After doing all our studies and analysis of the dataset we will try to summarize the obtained conclusions in three different parts, each one related with its chapter.

About the topology of our network we can conclude that:

- Both indegree and outdegree distribution follows a clear power law distribution.

- Users with larger outdegree than indegree have a fast decay in the histogram graph. This behaviour is due to new users who need to start friend request. A possible solution will be a mechanism of notifications with information about new users for old users with higher indegree. Then, new users will have an easier friend request start.

- Most of users have friendship relations approved by both sides.

- Strong linear dependence between indegree and outdegree exists for each user.

- Analyzing both graphs, the taste graph which is a subset of the larger network and is contained in the social graph is representative for the whole network in terms of need degree and link symmetry. Then, we can carry out our activity measurements only for the taste graph and we will not have problems of memory.

Referring to the crawling process and the power law generator we can sum up telling that:

- Any existing power law generator carry out two basic conditions to create a network similar to the Last.fm friend network: a generator with directed links and different exponents for the indegree and the outdegree power law exponents (1), and secondly, the condition based on the correlation between the indegree and the outdegree for each user (2).

- Anomalies in our graph were due to the stopping moment of the crawling process. The crawling based on the breath first method was stopped randomly, and the collection of some users' data was not finished on that moment. A possible solution will be to include a condition in the algorithm that only could stop the process when all the data collection was at least complete for a number of users, and those that were not complete, should be eliminate.

About the relation between the topology and user's activity we can state that:

- Extreme users interested in the listening activity are not interested in the friendship network. The opposite case is also confirmed, with users with higher indegree and fewer listening activity.

- Users who listen to popular artists have fewer indegree than the overall mean because they do not need to share their interests with other users and they are not interested in finding new or obscure artists from other profiles. Consequently, they are less interested in the social activity of the network and have fewer friends.

- Only 65% of our dataset is used to tag their songs.

- Users with more tagging activity have larger indegree and then, also more social activity in the network. This means that they are exploiting the social network site and helping other users to find music.

- Users that have a friendship relation have more tags in common. This conclusion is due to the fact that friends have common vocabulary and tend to have similar music tastes.

- Consequently, not all users take advantage of the site's social network functions. For some, the ability to stream music, keep track of their personal music charts, and receive music recommendations are their sole motivations for using the site. For others, the relationships between users are their motivation.

Chapter 8. Future Research

The aim of this chapter is to detail two main aspects that can be taken into account in the future. With this information another researcher interested in this project could continue developing it.

Our first idea was explained in section 4.3. We observed in Figures 8 and 9 that users with outdegree larger than indegree had a fast decay. This behaviour could be due to new users who need to start friend request. Our new idea consists on creating a small model based on notifications for old users with information about new users. With this model new users would have easier friend request start.

Our second idea was explained in section 4.5. There, we compared our results for the social graph of around 300.000 users with the taste graph of around 120.000. We can observe that results were really similar and with the taste graph we had sufficient data if we analyze the node degree and symmetry. It is important to remember that the taste graph is a subset of the social graph. Consequently, other possible work for the future will be to study which would be the break point of sufficient users for our analysis, searching the moment when results would not be representative. It would be done crawling samples from the same user with the first breath method and stopping the process in different times. Then, we would be able to determinate the size of our data which would be representative for the whole network, and with it, study and carry out all the measurements.

Finally, we obtained an idea from section 6.3.3 to improve the social network site. Friends have more tags in common and consequently, they use similar vocabulary and have similar tastes. If friends have similar interest they could be exploited to predict recommendations for a user. Then, if a user is searching for new music and we show him what his friends are listening, we are probably recommending him useful information. Also the idea of targeted advertising for users depending on their music interests would be interesting from investment and system design point of view.

Chapter 9. Acknowledgements

This MSc work has been carried out at the Network Architectures and Services (NAS) Group, Faculty of Electrical Engineering, Mathematics and Computer Science at the Delft University of Technology.

Firstly, I would like to thank to my director Christian Doerr for following the project step by step every day and helping me to achieve the goals.

I would like to thank my supervisors Siyu Tang for helping me with the graph theory and Maarten Clements for giving me support with Matlab questions. Also to Javier Martin for helping me with C++ programming problems.

Finally I would like to thank my family and friends for supporting me always.

Chapter 10. Bibliography

[1] Boyd, d. m., & Ellison, N. B.*: Social network sites: Definition, history, and scholarship*. Journal of Computer-Mediated Communication, 13(1), article 11.(2007).

[2] Travers, Jeffrey & Stanley Milgram. *An Experimental Study of the Small World Problem*. Sociometry, Vol. 32, No. 4, pp. 425-443.(1969)

[3] Richard Jones.: *Last.fm Radio Announcement.* (Tuesday, 24th March 2009)

[4] Richard Jones.: *Last.fm Radio Announcement.* (Monday, 30th March 2009)

[5] Baym, Zhang, Kunkel, Ledbetter, & Lin.: *Relational Quality and Media Use in Interpersonal Relationships* (2007)

[6] Hanneman, Robert A. and Mark Riddle.: *Introduction to social network methods*. (2005)

[7] I.B.Theisler.: *Topological Characteristics for measuring real-world networks.* (2007)

[8] Barabasi,A. and Albert,R.: *Emergence of scaling in random networks*, Science 2866, pp. 509-512 (1999)

[9] Newman M. E. J.: *Power laws, Pareto distributions and Zipf's law*.(May 2006)

[10] J. Martin Hernandez, T. Kleiberg, H. Wang , P. Van Mieghem.: *A Qualitative Comparison of Power Law Generators*, Delft University of Technology. (2006)

[11] *Igraph Library*. http://igraph.sourceforge.net/

[12] *Boost Library*. http://www.boost.org/

[13] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattacharje : *Measurement and Analysis of Online Social Network*. (2007)

[14]http://www.last.fm/charts/artist?charttype=weekly&subtype=artist&range=1212321600-1212926400

[15]http://www.last.fm/charts/hypeartist?charttype=weekly&subtype=hypeartist&range=1212321600-1212926400

[16] W. Aiello, F. Chung, L. Lu. *A Random Graph Model for Massive Graphs*. (2000).

[17] *Brite*. http://www.cs.bu.edu/brite/

[18] Cardillo, Scellato and Latora. *A topological analysis of scientific coauthorship networks*.(2006).

[19] Bu and Towsley. *On Distinguishing between internet power law topology generators*. IEEE pp 638-647 (2002).

[20] R. Albert, and A. Barabasi. *Topology of evolving network: Local Events and universally*. (2000)

[21] S.L. Hakimi. *On the realizability of a set of integers as degrees of the vertices of a graph*. SIAM J. Appl. Math, 10, 1962.

[22]Takao Asano. An *O(n log log n) Time algorithm for Constructing a Graph of Maximun Connectivity with Prescribed degrees*. Journal of Computer and System Sciences, 51, 503-510 (1995).