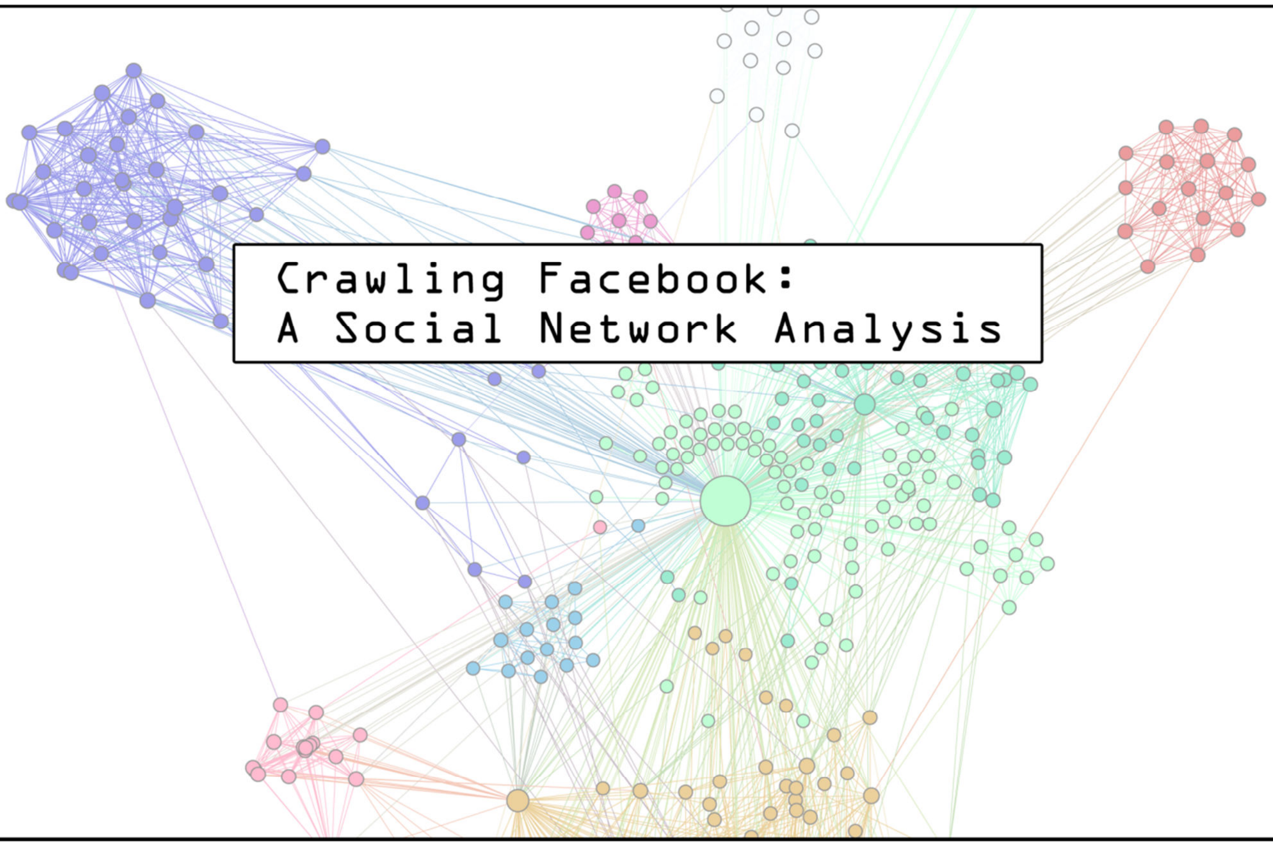
MScIS
MSc in Information Systems
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS



Crawling Facebook: A Social Network Analysis

Theodore Papageorgiou *M4090017*

Master Thesis

Supervisor: Prof. Vazirgiannis Michalis

External Reviewer: Prof. Stamoulis Georgios

ATHENS SEPTEMBER 2011

ABSTRACT

Online Social Networks (OSN) play an integral role in our everyday life, affecting the social life and activity of people in various ways. Social Networking sites have hundreds of millions of registered users who use these sites to share thoughts, experiences, photographs, meet new people, contact long-lost friends and family members, find jobs, spread information, and more

The idea of social networks, and that social phenomena can be explained when we surpass the properties of individuals and examine their personal and social ties, has been around for over a century.

Social Networks play a critical role in the social, economic, health, educational aspects of our life and behavior in general. Their structure affects the way information flows amongst people, the way diseases spread, our purchase choices, the decisions we make and the way our society evolves.

In this Thesis we perform a study that includes crawling the most popular online social network site "*Facebook*" and performing a proof-of-concept Social Network Analysis. We describe the collection process of the crawlers implemented in python. Moreover we provide graph visualization and study several graph metrics with the help of Gephi, an open source program for visualizing and analyzing large graphs. We provide metrics and analyze network graph properties such as degree distribution, centrality measures, and community detection, among others.

From our extracted anonymized data we choose to further analyze users' likes in conjunction with their relationships and provide basic statistics and analysis. We analyze the community detection mechanism and raise the question if community unfolding results can be reproduced and/ or improved or if we take into consideration the users common preferences (likes).

Keywords: crawler, data mining, facebook, social network analysis, graph analysis

ACKNOWLEDMENTS

I would like to thank my supervisor Professor Michalis Vazirgiannis and PhD Student Mr. Nikos Salamanos, without whose help, guidance and support this Master Thesis would not be realized.

CONTENTS

1. Introduction

Online Social Networks (OSN) play an integral role in our everyday life, affecting the social life and activity of people in various ways. Social Networking sites have hundreds of millions of registered users who use these sites to share thoughts, experiences, photographs, meet new people, contact long-lost friends and family members, find jobs, spread information, and more. *Facebook (FB)* is the world's largest online social network, with 750 million users worldwide as of July 2011, with 50% of the active users logging on to Facebook in any given day.[1] Online Social Networks do not differ much from earlier Social Networks aside from the mechanism used by the members to communicate with each other. In the online world, communication is facilitated with Web Technologies, whereas on earlier Social Networks, communication encompassed face-to-face interaction.

The idea of social networks, and that social phenomena can be explained when we surpass the properties of individuals and examine their personal and social ties, has been around for over a century. In the late 1800s the work of sociologists such as Ferdinand Tönnies[1] and David Émile Durkheim[2] argue about the associations between members of communities and collectives. In the 1960s-1970s, significant work by numerous scholars in sociology departments such as Linton Freeman, Harrison White, S.D. Berkowitz Mark Granovetter, Peter Marsden, Nicholas Mullins, Barry Wellman, Anatol Rapoport to name but a few, had elaborated and popularized social network analysis.

Social Networks play a critical role in the social, economic, health, educational aspects of our life and behavior in general. Their structure affects the way information flows amongst people, the way diseases spread, our purchase choices, the decisions we make and the way our society evolves. The interest in Online Social Network Analysis has been growing massively in recent years. Social Network Analysis has been a key technique for sociologists along with anthropologists, psychologists, biologists, economists, and statisticians constituting it an interdisciplinary research area.

---

[1] http://www.facebook.com/press/info.php?statistics

1.1 Network Analysis

We will provide the fundamentals of how networks are represented, measured and characterized. Some basic concepts and definitions that are fundamental in Network Analysis are provided in the following paragraphs.

Individuals belonging to a network are being referred as ***Nodes***. They are often referred as *Vertices, Actors* and *Agents* amongst others. A set of Nodes participating in a network of relationships is symbolized as $N = \{1,\ldots,n\}$. Depending on the context of the analysis, and the nature of the Network in question, nodes may vary from people, countries, webpages and ontologies to molecules and proteins.

The nodes of the network can be either connected or not. The relationship ties between nodes are referred as ***edges****, links,* or *connections*. The edges can be either mutual or not. We can imagine a mutual or undirectional edge when representing friendship between two persons/ nodes, where ties are necessarily reciprocal, and a directional tie, in situations where a node can link to another without getting consent, i.e. a webpage linking to another.

Networks are represented by graphs. A ***graph G (V, e)*** consists of a set of nodes $V = \{1,\ldots,n\}$ and a real-valued $n \times n$ matrix v, where v_{ij} represents the (possibly weighted and/or directed) relation between i and j. This matrix is called the adjacency matrix, listing which nodes are linked (are adjacent) to each other.

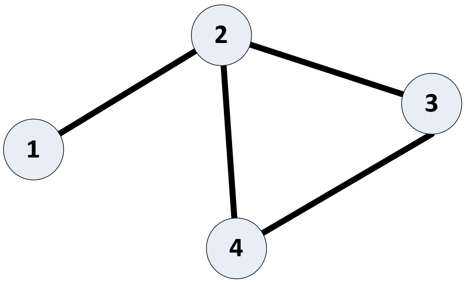


Figure 1: A network with four (4) nodes and four (4) links

A graph can be referred to as a ***weighted graph***, if the entries of e can take more than two values and we can cognize the intensity on the level of these relationships.

A network is called ***undirected*** when $e_{ij} = e_{ji}$ for all nodes i and j. A network is ***directed*** when $e_{ij} \neq e_{ji}$ for all nodes i and j. Directed graphs are often referred to as ***digraphs***.

If $V = \{1, 2, 3, 4\}$ then,

$$e = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

is the undirected/ unweighted network illustrated on Figure 1.1.

1.2 Characteristics of Networks

Degree d_i of a node n_i, is the number of neighbors (nodes) adjacent to n_i. The degree of *node #2*, from Figure 1.1. is $d_2 = 3$. Nodes with zero degree are called *isolated nodes*. Degree level in directed graphs is further divided to ***In-Degree*** and ***Out-Degree***, referring to the number of directed incoming links towards a node and outgoing to another node.

In real-world networks not all nodes have the same degree. Also the degree of a *node n* and *Average Degree* of a *graph G*, is a dynamic value changing through time.

A ***path*** is a sequence of nodes such that from each node in the path there is an edge to the next node in the sequence. A ***connected*** graph is a graph where there exists a path between any pair of nodes. A maximal connected subgraph of a graph is called ***component***.

Distribution P(d) is a function that describes the probability of a random node having a certain degree d and is called degree distribution, referring to the spread of nodes' degrees in a network. There are some well-known degree distributions such as *Poisson*, *power-law* and *exponential* distribution. Networks with ***Power-Law distribution*** are called ***Scale-Free***.

The most common encountered in practice distribution is the *normal (or Gaussian) distribution*. For example the average male height in Greece is 1.781 m (5 ft 10 in)[2] while the rest of measurements are symmetrically distributed around their mean, yielding a "bell" curve plot.

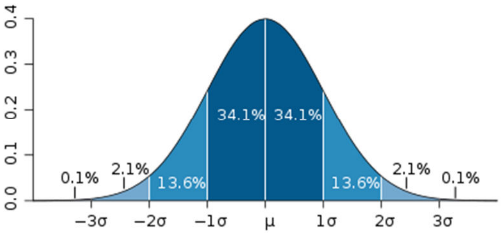


Figure 2: Normal Distribution[3]

Nodes' degree in real-world, large scale social networks often follow a *power law distribution*[3]. Networks whose degree distribution follows a Power Law are Scale-Free Networks. In *Scale-Free Networks* we observe nodes with a high degree that greatly exceeds the average. These highest-degree nodes are called ***hubs***, and are considered to play a more significant role in a network.

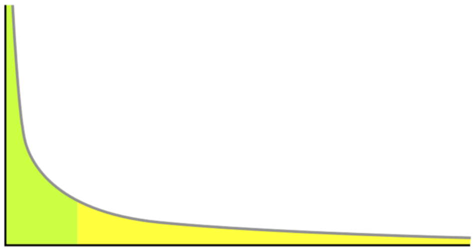


Figure 3: Power Law Distribution[4]

A power law probability or frequency distribution of a given degree can be expressed as

$$P(d) = cd^{-\gamma}$$

where $c > 0$ and $\gamma > 1$ are parameters of the distribution, and hence the term *power-law*. *The scale-free aspect refers to the fact that if we consider the*

---

[2] http://www.elkede.gr/images/EthnikiSomatrometrikhEreuna-page2.pdf
[3] Figure used under the Creative Commons Attribution 2.5 Generic license
[4] Picture by Hay Kranen (http://www.haykranen.nl/)

probability of a degree d and compare that to a degree d', then the ratio of $P(d)/P(d') = (d/d')^{-\gamma}$*. Now suppose that we double the size of each of these degrees. We find that* $P(2d)/P(2d') = (d/d')^{-\gamma}$*. It is easy to see that rescaling d and d' by any factor will still give us this same ratio of probabilities, and hence the relative probabilities of different degrees just depends on their ratio and not on their absolute size. This explains the term scale-free [M.* Jackson*]*[4].

Clustering coefficient, measures the number of edges between neighboring nodes of a node. Two versions of Clustering coefficient exist. The *Global* that provides an overall indication of the network's clustering and the *Local* that provides an indication of the *embeddedness* of single nodes.

The ***Local Clustering Coefficient*** [Watts and Strogatz][5], of a node v_i is given by the ratio of the existing edges e_i from that node to its neighbors, and the total number of edges that could exist between them $k_i(k_i - 1)$ also known as *Neighborhood*.

$$\mathrm{C}(vi) = \frac{2ei}{\mathrm{ki}(\mathrm{ki} - 1)} \quad \text{, for } \textit{undirected graphs}$$

$$\mathrm{C}(vi) = \frac{ei}{\mathrm{ki}(\mathrm{ki} - 1)} \quad \text{, for } \textit{directed graphs}$$

The ***Global Clustering Coefficient*** [Luce and Perry][6] of a node v_i is based on *triplets*, where a ***triplet*** is three nodes connected with either two or three undirected ties. A triplet with three ties is called a *closed triplet*, whereas a triplet with only two ties is called an *open triplet*. A triangle consists of three closed triplets, one centered on each of the nodes. From the above two definitions we have:

$$\mathrm{C}(vi) = \frac{\text{3 x number of triangles in the graph}}{\text{number of connected triples in the graph}}$$

Centrality. In Social Network Analysis it is important to discover the relative importance of nodes and identify the ones that have better access to information and are more capable of spreading it through the network. We will refer to 4 centrality measures: D*egree Centrality*, *Betweenness Centrality*, *Closeness Centrality*, and *Eigenvector Centrality*.

Degree Centrality is one of the simplest *Centrality* measures that shows us how well connected a node is. The degree centrality of a node *i* of graph with *n* nodes is calculated simply by:

$$\mathrm{CD(i)} = \frac{\mathrm{degree(i)}}{(\mathrm{n} - 1)}$$

ranging from 0 to 1 and expressing how well a node is connected, in relation to its direct connections. However Degree Centrality is not sufficient to express the importance of a node concerning its position in the network. A node might have relatively few connections and also lie in a more critical and influential location in the network.

Betweenness Centrality measures the number of times a node lies in the paths of a graph. First proposed by Freeman[7], *Betweenness* $C_B(i)$ of a node *i*, is calculated as follows

If $P_i(hj)$ is the number of *shortest paths (geodesics)* between nodes *h* and *j* that *i* lies between and *P(hj)* is the total number of shortest paths between *h* and *j,* then by calculating the ratio of $P_i(hj)/ P(hj)$ we can find the importance of node *i* in the relationship of *h* and *j* nodes. Values close to 1 indicate that node *i* is highly important, whereas values close to 0 indicate that node *i* is of little importance for them.

If we normalize by dividing through the number of pairs of nodes not including *i*, which is $(n - 1)(n - 2)$ for directed graphs and $(n - 1)(n - 2) / 2$ for undirected graphs we get

$$CB(i) = \sum_{h\neq j:i\notin\{h,j\}} \frac{Pi(hj)/\, P(hj)}{(n - 1)(n - 2)} \quad \text{, for directed graphs}$$

$$CB(i) = \sum_{h\neq j:i\notin\{h,j\}} \frac{Pi(hj)/\, P(hj)}{(n - 1)(n - 2)\, /\, 2} \quad \text{, for undirected graphs}$$

Closeness Centrality measures how close a node is to each other node in the graph. The closeness $C_C(i)$ of a node *i* the inverse of the average distance between *i* and any other node

$$CC(i) = \frac{(n - 1)}{\sum_{j \neq i} l(i,j)}$$

where l(i, j) is the number of links in the shortest path between *i* and *j*. Smaller values of Closeness Centrality means the greater the distance of a node with any other nodes, thus less chances of receiving information.

Eigenvector Centrality is a measure of the importance of a node in the network based on how influential and important are its neighbours. Google's Page Rank[8] algorithm is based on the Eigenvector Centrality. Eigenvector Centrality assigns relative scores to the nodes of a network defining their "popularity", with connections to high-scoring nodes contributing more to the score of the node under investigation, and connections to low-scoring nodes contributing less.

If $G(V, E)$ is a graph, consisting of nodes V and edges E and A is the adjacency matrix of G . If node i is linked with node j, then $a_{ij} = 1$, and if not $a_{ij} = 0$.

The *Centrality* score for node i will be proportional to the sum of all Centrality scores of the nodes to which it is connected. Therefore

$$x_i = \frac{1}{\lambda} \sum_{i=1}^{N} \alpha_{ij} \, x_j$$

where N is the total number of nodes and λ is a constant which is the largest eigenvalue of A

$$Ax = \lambda x$$

1.3 Research on Social Network Analysis

Social network analysis has provided substantive contributions in the areas of sociology, economics, interorganizational relations, social influence, and epidemiology.

In epidemiology it has helped understand how patterns of human contact aid or inhibit the spread of diseases. In a recent example, researchers combined two tools to get a clearer picture of a tuberculosis outbreak: social network analysis, which has become increasingly common in tracking infectious diseases in the past decade, and whole-genome sequencing. "*Public health agencies are now able to harness the power of genome sequencing, which, when combined with the detailed clinical and epidemiological data we have access to, allows us to reconstruct outbreaks and really understand how a pathogen moves through a population*". [Jennifer Gardy][9]

Social Network Analysis is being used to study the dynamic spread of ideas, concepts and trends over the Internet, providing insight on how people interact, and the implications of how they are associated. *Viral marketing can be thought of as a diffusion of information about the product and its adoption over the network. Primarily in social sciences there is a long history of the research on the influence of social networks on innovation and product diffusion* [Leskovec, Adamic, Huberman][10]. Diffusion of innovations theory [Everett M. Rogers][11] seeks to explain how people are influenced by and influence the spread of ideas, and decision making. Rogers refers to the agents of a network as opinion leaders and followers. Social Network Analysis provides the tools to observe the effectiveness of person-to person, word-of-mouth advertising, thus making it a decisive tool for marketing campaigns. The diffusion model proposed by Bass in 1969[12], that contributed to Rogers model, predicts the number of people who will adopt an innovation over time. It does not explicitly account for the structure of the social network but rather it assumes that the rate of adoption is a function of the current proportion of the population who have already adopted it (purchased a product).

Social Network Analysis can be used as an effective tool applied for surveillance against criminal threats or at the prosecution of criminal activities. SNA has a long history of application to evidence mapping in both fraud and criminal conspiracy cases. An attempt to uncloak terrorists' networks after the tragic events of September 11th 2001 by Valdis E. Krebs[13] can be found in INSNA website[5]. Additionally the Information Awareness Office (IAO) that was established by the Defense Advanced Research Projects Agency (DARPA) in January 2002, attempted to create an enormous database for the personal information of everyone in the United States, including personal e-mails, social networks, credit card records, phone calls, medical records amongst other data, in order to be analyzed and identify potential threats. The IAO faced the public criticism and was soon defunded by the Congress in 2003, however several parallel projects continue to run until this day.

Stanley Milgram's famous study of the small-world phenomenon[14] in 1967, demonstrates that in large graphs numerous short paths exist and that information is able to find its way and get distributed across these paths, even if the map of such graph is unknown. Milgram asked random individuals living in the cities of Omaha, Kansas, Nebraska and Wichita to try to forward a letter to a designated target in the cities of Boston and Massachusetts. If the "starter" individuals knew the "target" recipient, they would send the letter to them. If not, they had to forward the letter to a single acquaintance that he or she knew on a first-name basis and the procedure would continue until the letters arrived at the target. Soon, letters began to arrive at their targets in a few as one or two hops, while others followed a path of nine to ten hops. Many of the letters failed to reach their destination as people refused to pass the letters forward, but for the letters that eventually did reach the target, it was estimated that on an average, 5.5 to 6 hops were required. This phenomenon was widely addressed as the "Six Degrees of Separation", and highlighted that the world is "smaller" than people thought and that on average only 6 hops were between two random individuals in the US. In 2007, Jure Lescovec and Eric Horvitz conducted an experiment in Microsoft Research facilities[15], using data captured from the Microsoft Instant

---

[5] http://www.insna.org/pubs/connections/mindex.html

Messaging system. They investigated the “Six Degrees of Separation” claim examining 30 billion conversations among 240 million people and found that the average path length among Messenger users is 6.6.

Dunbar’s number is a theoretical cognitive limit of the number of people a person can maintain stable social relationships with. Robin Dunbar suggested in a 1992 article that the typical size of an egocentric network is constrained to about 150 members (the number lies between 100 and 230) due to possible limits in the capacity of the human communication channel or else the neocortical processing capacity. Dunbar's number has since become a constant taken in regard by sociologists, anthropologists, statisticians, psychologists and other.

1.4 Overview of the Present Analysis

Motivated by the rising interest in Online Social Networks Analysis and Facebook’s amazing growth, we decided to perform a proof-of-concept study that included crawling a real online social network and analyzing the collected data. Albeit a number of datasets already existed in the academia, we preferred to go through the tentative procedure of collecting one of our own. Prior to this attempt, no experience on crawlers and social network analysis was possessed, so this procedure included a lot of trial, analysis and creative problem solving efforts. The data (anonymized dataset available upon request) was collected by designing and running focused crawlers, implemented in Python and storing the data locally on a database for further processing. The initial strategy needed to change numerous times, through this trial and error period, as Facebook’s design sets many obstacles and limitations on the data publicly available.

In the following section, we describe the methodology and the process followed that we designed in order to collect the Facebook data.

1.4.1 Data Collection Process

The designed Python scripts log in to Facebook with an existing User Account and can access all kind of personal information publicly available to the logged in user such as friend list, gender, age, locale, current location, hometown, school, wall posts, interactions, page likes, and more.

Crawling Facebook was not an easy task, since the platform raises limitations due to the restrictive data access policies. There are two sources the scripts are able to collect data:

(i) the Facebook Graph API which provides authorized third-party developers a simple, consistent view of the Facebook's social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags)[6].

(ii) scraping raw HTML source code. FB's Graph API, while being a useful tool, provides in purpose limited information in proportion to what is publicly available while browsing through user profiles. For this reason, a script was designed that simulated browsing patterns of a registered user. Facebook's architecture makes it difficult to access and scrape useful information from each user profile, since everything is displayed dynamically with asynchronous feeds on request. On top of it, Facebook engineers continuously perform under-the-hood changes, requiring respective modifications to our script. In order to discover the mechanism of these asynchronous feeds an intercepting proxy server was used, to sniff HTTP traffic passing from our web browser to Facebook.

1.4.2 Analyzing crawling process

The scripts log with an existing Facebook account (the author's account) and collect the friend list of that specific profile. We then iterate through each fetched friend id and collect their friend lists, repeating for as many levels we choose. We can extract all kind of personal information publicly shown such as gender, age,

---

[6] http://developers.facebook.com/docs/reference/api/

locale, current location, hometown, school, wall posts, interactions, page likes etc. There were days that our scripts run on a 24/7 basis, sending thousand requests to Facebook's servers per crawling session. In contrary to other documented attempts to crawl Facebook, our scripts never alerted Facebook's Security mechanisms hence our account was never suspended.

The scripts, written in Python make use of these libraries amongst others:

urllib, urllib2 modules necessary for the http connection

MySQLdb module necessary to read/write data in the MySQL database

re module necessary to parse and scrape data from the raw html source code using regular expressions

simplejson module necessary to parse JSON data from Facebook's graph API response

1.4.2.1 Fetching Friend Lists

Facebook's continuous development and layout changing made this task quite difficult as crawling can be a time consuming procedure and a change would mean that one would need to troubleshoot and re-engineer the code and find different methods to crawl the profiles. When viewing a user profile and its friends, Facebook displays friends in a dynamic way, by feeding 60 friends at a time and adding them to the list with multiple asynchronous requests. In order to discover the source of these feeds and try to imitate this procedure an intercepting proxy server was used on our system, in order to sniff HTTP traffic passing from the web browser to Facebook's server. It was found out that when calling the url below:

http://www.facebook.com/ajax/browser/list/friends/all/?uid=**X**&offset=**Y**&dual=1&__a=1

Where **X** is the user's profile ID we want to gather their friend list from and **Y** is the offset number of the friends in lots of 60, so beginning with **Y** = 0, then **Y** =

60, then **Y =** 180 the first 180 friends of that person were gathered, and by serially adding 60 the whole list was collected.

By repeating this procedure for all ids we managed to gather all friends of the nodes under examination.

1.4.2.2 Fetching User Information

In order to fetch User information such as name, username, birthday, hometown, location, work information, education, gender, locale, languages, etc per crawled ID there was no need to visit each user's profile and scrape the data out of it. Facebook's Graph API provides an easy interface to view all this information in an instance by visiting the url below:

https://graph.facebook.com/**X**?access_token=**Y**

Where **X** is the users profile ID we want to gather the information from and **Y** is the Access Token using the OAuth 2.0 protocol for authentication and authorization someone can scrape from Facebook's GRAPH API page.

http://developers.facebook.com/docs/reference/api/

Then it is just a matter of fetching the desired information through the JSON response of the above link.

1.4.2.3 Fetching Users' Likes

User's likes can be fetched in two different ways. Facebook's Graph API provides a way to view all the pages a user has liked, plus the exact timestamp this user had liked each page.

But it seems that you are only authorized to view your friends' likes. If you try viewing your friends' friends likes or people you are not directly related with, it fails most of the time. So for profiles outside our friendship range a different solution is applied.

Once again this required careful observation and reverse engineering of the way each user's profile page is constructed when you view their likes. By sniffing HTTP requests we found out that each user's likes that are presented in their profile under categorization are feeded from 10 different sources in accordance to their categories. These categories and source links are:

Favorite teams:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**129476497102318**&profile_id=**X**&offset=0&__a=1

Favorite athletes:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**162363193777361**&profile_id=**X**&offset=0&__a=1

Music:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**13001**&profile_id=**X**&offset=0&__a=1

Books:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**13002**&profile_id=**X**&offset=0&__a=1

Movies:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**13005**&profile_id=**X**&offset=0&__a=1

TV:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**13006**&profile_id=**X**&offset=0&__a=1

Games:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**164382953603504**&profile_id=**X**&offset=0&__a=1

Other Pages:

http://www.facebook.com/ajax/profile/information.php?profile_id=**X**&section_ids%5B0%5D=9999&meta_section=9999&content_id=otherid_4d738352539219334062543&__a=1

Activities-Interests:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**1002**&profile_id=**X**&offset=0&__a=1

Sports:

http://www.facebook.com/ajax/profile/show_more.php?section_id=**100000450914378**&profile_id=**X**&offset=0&__a=1

where **X** is the User's Profile ID

Combining the results of these 10 different sources we were able to gather all the pages a user has liked. A downfall of this method compared with the Facebook Likes Graph API, is that we are missing the timestamp of each performed like, but still being able to gather each user's preferences is valuable. One thing we noticed with this method is that occasionally we gathered pages that were not shown in graph API. These were pages under "Education and Work" category and are Institutions or Workplaces that one of your friends has tagged you as being schoolmates or workmates. This is considered as valid, added value information as it is information published in each user's profile.

1.4.2.4 Fetching Pages' Information

The pages being fetched in the previous paragraph were just Page IDs. Until this moment no information is held regarding what these pages represent. Using

Facebook's Page Graph API, it is easy to find the Page's name, how many times it has been liked in total and under which category this page is classified, among other information. The Python script checked the information for every added page and recorded it in the database.

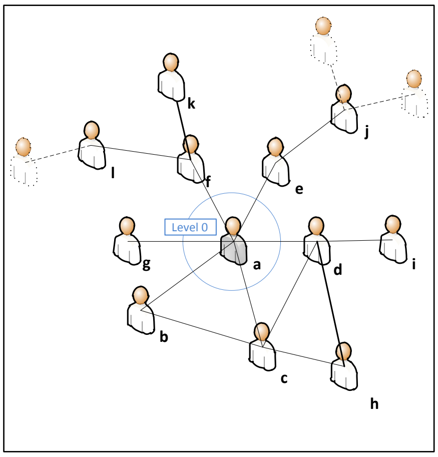
1.4.3 Description of the Collected Data

Over six gigabytes of user data split into various datasets was collected during the period of writing this thesis. For this study, three datasets are presented, which are summarized in Table 1. The main point of interest in all these datasets was collecting friend lists, likes, and demographics per crawled node. The crawling method that was chosen to be used for this particular study is the most widely graph traversal technique, *Breadth-First-Search (BFS)*. BFS has been used extensively for sampling Social Networks and for this study we performed a neighborhood-constrained BFS, with an outer limit of the friends of friends, of the starting node. BFS has shown to be biased towards high degree nodes[16], but for the purposes of this analysis we find the method effective.

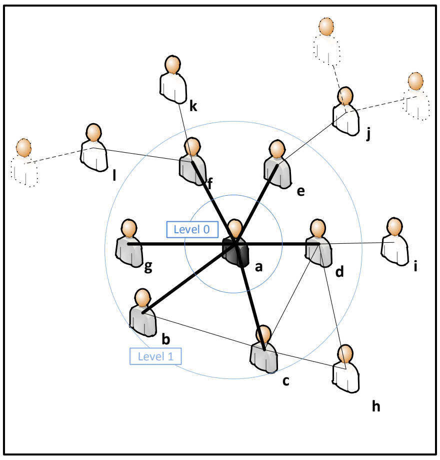
1.4.3.1 Overview of the Breadth-First Search Traversal Algorithm

Breadth-first Search Algorithm begins at a given source node *a* of a graph *G*, which is at *level 0*. In the first stage, the crawler visits all neighboring nodes that are at the distance of one edge away. These discovered nodes are labeled as visited and belong now to *level 1*. Node *a*, is labeled as explored. In the second stage, we will explore all discovered nodes of *level 1*, visiting all the new nodes we can reach at the distance of two edges away from the source node *a*. These newly discovered nodes *(visited)*, which are adjacent to *level 1's* nodes and not previously discovered, belong to *level 2*. All *level 1's* nodes are now explored. These steps are repeated until every node of *graph G* has been visited - or in our occasion - when we visit the nodes of *level 2*. For our purposes we perform an incomplete BFS, as we are only interested in the adjacent neighbors of our source node *a*. We will refer to the node at *level 0* as *source node*, the nodes ta *level 1* as "*friends*" (of the source node) and the nodes at *level 2* as "*friends of friends*", following Facebook's nomenclature.

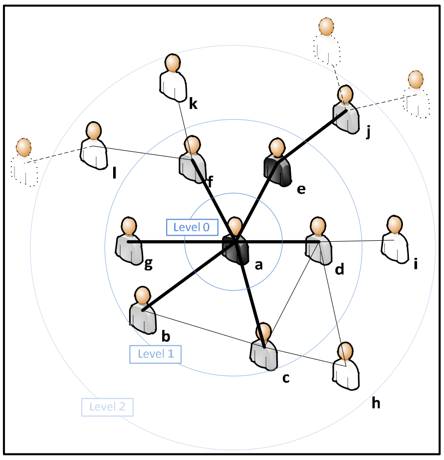
The following figures illustrate the progress of the breadth-first search crawling on Facebook's undirected graph.



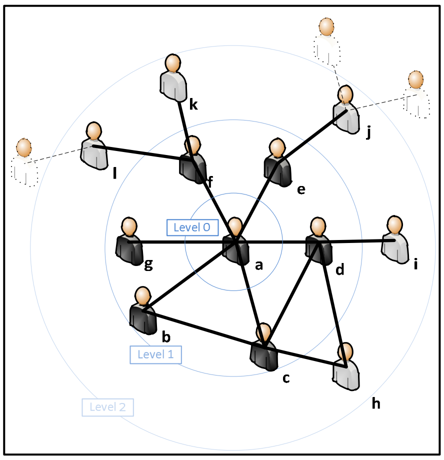
Given the Facebook graph, we visit the selected source *node a*, that we assign at *level 0*.



We explore *node a*, and discover all adjacent nodes, one step far from our source node. The darker lines, indicate the discovered edges and the grey figures the *discovered* (*visited*) nodes. These nodes belong to *level 1. Node a*, is colored in black, meaning it's an *explored* node.



We then repeat the procedure for *node e* of the *level 2*, discovering *node j* that belongs to *level 3. Node e*, is marked with black, meaning that it is an *explored* node and *node j* is marked with grey, meaning it is a visited node. This procedure will be repeated for all nodes of *level 1*, as illustrated on the following figure.



We incrementally explore all nodes of *level 1*, discovering *nodes I, k, j, I, h* that are labeled as *visited* (grey color) and belong to *level 2*. We stop at this point and ignore nodes of next levels.

1.4.4 Datasets - Collection Process

Starting from source *node a* (author's profile), we crawl the Facebook graph using the Breadth-First-Search technique, with depth of one level. The newly discovered nodes and edges plus *node a* constitute graph G_f (n_f, e_f) or node *a*'s "*Friends*". We then continue with our Breadth-First-Search mechanism to the next level, until we discover all nodes of *level 2* or node *a*'s "*Friend of Friends*". The total discovered edges and nodes constitute *graph G(N,E).* Given our two graphs we collect the following datasets.

1.4.4.1 Dataset I (Graph Preferences)

In *Dataset I* we explore all 100.390 discovered nodes of *Graph G* (*Node a, Friends and Friend of Friends;* Instance taken on 18/05/11) and collect their preferences (*likes*). A total of 790.447 *pages* were discovered with 5.858.958 *likes* linking to them. The process to obtain the likes took 5 days and the process to obtain the information of such an amount of pages(name, category, etc) took 4 weeks.

1.4.4.2 Dataset II (Uniform Evolution Sample)

From *graph G(n,e)* taken on 18/05/11 we select five (5) hundred random nodes. For these five hundred nodes, we collect on a daily basis their friends and their preferences (likes). We repeat for ten (10) consecutive days, resulting on a dataset that we can examine the evolution in growth and preferences (likes).

1.4.4.3 Dataset III (Friends Evolution Sample)

For *Dataset III*, and on a daily basis we crawl the Facebook graph using the Breadth-First-Search technique starting from the same *node a*. We collect node *a*'s friends and friends of friends. We then collect the preferences (likes) of node *a* and it's *friends*. We repeat for twenty (20) consecutive days, resulting on a dataset that we can examine the evolution in growth and preferences (likes) and compare it with *Dataset's II*, random nodes.

Dataset	Elements	Period
I (Graph Preferences)	Likes, Nodes & Page Info	09/05/11 – 14/05/11
II (Uniform Evolution Sample)	Friend lists, Likes, Nodes & Page Info	18/05/11 – 27/05/11
III (Friends Evolution Sample)	Friend lists, Likes, Nodes & Page Info	08/05/11 – 27/05/11

Table 1: Datasets

1.4.5 Information collected per discovered user

When applicable and/ or published:

- **Friend list:** friendship is always mutual thus leading to undirected edges
- **UserID:** each user is uniquely defined by a 32-bit userID. This userID is hashed in our database, in order to keep no reference with the real user.
- **Gender:** user's gender
- **Locale:** the selected UI language for Facebook Site
- **Birthday:** the Birthday Date the user has supplied
- **Location/ Hometown:** Information regarding user's hometown and current town
- **Pages:** User's preferences indicated by "Likes"

1.4.6 Information collected per discovered page

When Applicable:

- **PageID:** each page is uniquely defined by a 32-bit pageID
- **Name:** the name of the page
- **Link:** url to the Page
- **Category:** the category this page belongs to
- **Fan Count:** Total Likes this page has gathered at the time we crawled it.

The above information is illustrated on the figure below.

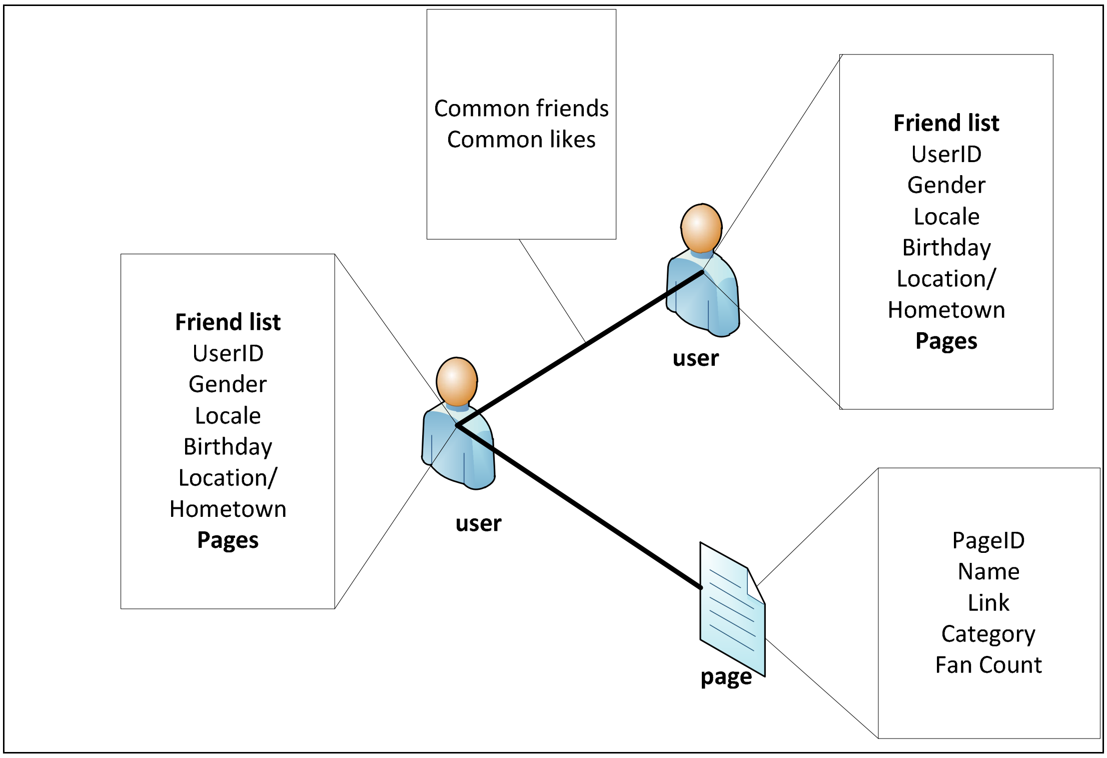


Figure 4: Information Collected per explored node

2 Data Analysis

In this section, the three datasets are analyzed, focusing on Dataset I. Dataset II and III are briefly presented with an intention to analyze furthermore in future reports.

2.1 Graph Demographics

We present matrixes and plots regarding distinctive characteristics of the collected users.

2.1.1 Locale

Facebook is currently available in over 70 languages[7]. Facebook locales follow ISO language and country codes respectively, concatenated by an underscore. The basic format is ''ll_CC'', where ''ll'' is a two-letter language code, and ''CC'' is a two-letter country code. For instance, 'en_US' represents US English.

#	Locale	Users
1	el_GR	42930
2	en_US	29657
3	en_GB	15630
4	sv_SE	3620
5	non_disclosed	1382
6	it_IT	1109
7	de_DE	1073
8	fr_FR	948
9	es_LA	754
10	nl_NL	414
11	pt_PT	362
12	es_ES	299
13	da_DK	259
14	fi_FI	254
15	lt_LT	163
16	ru_RU	156
17	tr_TR	144
18	bg_BG	129
19	sr_RS	113
24	en_PI	68
25	nb_NO	56
26	et_EE	53
27	id_ID	51
28	cs_CZ	48
29	ja_JP	48
30	hu_HU	47
31	ar_AR	30
32	hr_HR	29
33	zh_CN	25
34	ro_RO	22
35	sl_SI	20
36	fr_CA	17
37	sk_SK	14
38	zh_TW	13
39	nn_NO	12
40	nl_BE	11
41	is_IS	9
42	th_TH	9
47	ko_KR	6
48	vi_VN	6
49	gl_ES	4
50	uk_UA	4
51	cy_GB	3
52	en_IN	3
53	kk_KZ	3
54	af_ZA	2
55	zh_HK	2
56	be_BY	1
57	bs_BA	1
58	en_UD	1
59	eu_ES	1
60	fo_FO	1
61	hy_AM	1
62	km_KH	1
63	mk_MK	1
64	ms_MY	1
65	mt_MT	1

---

[7] http://developers.facebook.com/docs/internationalization/

20	pt_BR	95
21	sq_AL	83
22	ca_ES	80
23	pl_PL	80
43	fb_LT	7
44	he_IL	7
45	la_VA	7
46	lv_LV	7
66	qu_PE	1
67	te_IN	1
68	xh_ZA	1

Table 2: Locale Distribution

The extracted information from the crawled users' profiles regarding locale distribution is presented in a graph below.

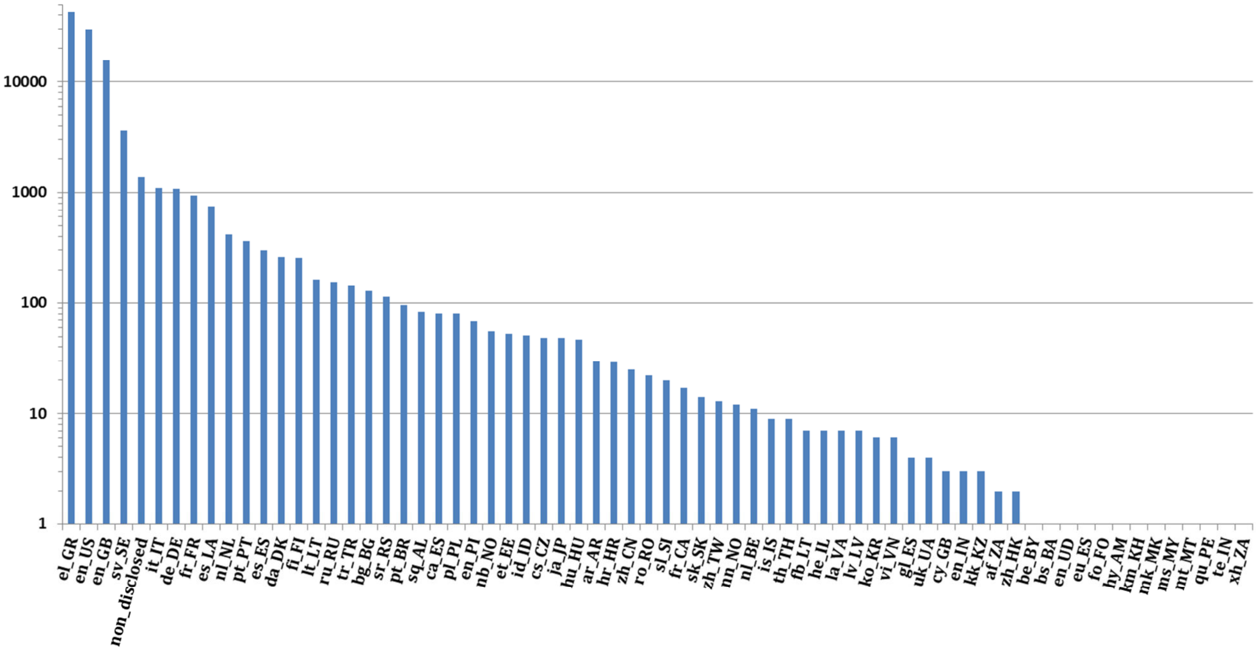


Figure 5: Locale Distribution

We observe that in the top 3 rankings we find: 1) Greek Language (42.930 users), 2) American English Language (29.657 users) and 3) British English (15.630 users). We expected to find Greek at the top position as the source Node (Author) we began our crawl with, is a native Greek citizen, thus the majority of his friends and acquaintances are of the same nationality.

2.1.2 Sex Ratio

By accessing the gender information provided on each user profile, we are able to analyze the sex ratio in our graph. We can observe the similarity with the official World Wide statistics provided by Cia's Fact Book[8].

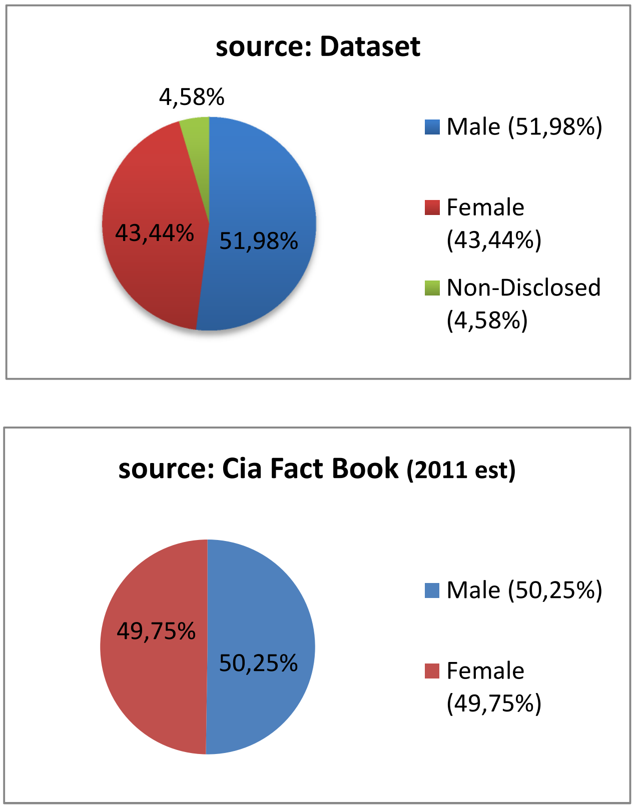


Figure 6: Sex Ratio

2.2 Dataset I (Graph Preferences) Analysis

We examine various aspects and statistics regarding the preferences of users belonging to the crawled graph.

---

[8] https://www.cia.gov/library/publications/the-world-factbook/geos/xx.html

2.2.1 Categories Popularity – Pages per Category

From the 100.390 profiles we crawled, 68.744 users had their liked pages published and with a total of 5.858.958 likes counted, we estimate that in average, a user has 85 likes. Facebook's Statistics[9] indicate that the *"Average user is connected to 80 community pages, groups and events."* From the above, we found 790.447 distinct pages, classified under 207 different categories. It is interesting to observe that the majority of pages are classified under the "Local Business" category. These are pages belonging to businesses promoting their brand and products, a widely used marketing technique nowadays. We present a table and a graph with the top 50 Categories and the number of pages per category.

#	Category	Pages
1	Local business	99193
2	Musician/band	74610
3	Website	51513
4	Community	49008
5	Company	45653
6	Public figure	44667
7	Uncategorized	31112
8	Interest	30972
9	Product/service	26425
10	Non-profit organization	22659
11	Movie	17390
12	Book	15224
13	Artist	13546
14	Athlete	13193
15	Actor/director	12799
16	Organization	10888
17	Games/toys	10644
18	Tv show	10185
19	Club	9442
20	Professional sports team	9430
21	Restaurant/cafe	8949
22	Author	7656
23	Music	7450
24	School	6156
25	Cause	5888
26	Hotel	5788
27	University	5560
28	Movie general	5076
29	Entertainment	4730
30	City	4507
31	Education	4294
32	Clothing	4276
33	Tv	3852
34	Politician	3793
35	Cars	3779
36	Bar	3615
37	Food/beverages	3380
38	Magazine	3263
39	Song	3152
40	Travel/leisure	3049
41	Media/news/publishing	2963
42	Games	2885
43	Radio station	2694
44	Personal blog	2694
45	News/media	2603
46	Arts/entertainment/nightlife	2544
47	Record label	2369
48	Album	2272
49	Health/beauty	2109
50	Shopping/retail	1902

Table 3: Top 50 Categories – Pages per Category

---

[9] http://www.facebook.com/press/info.php?statistics

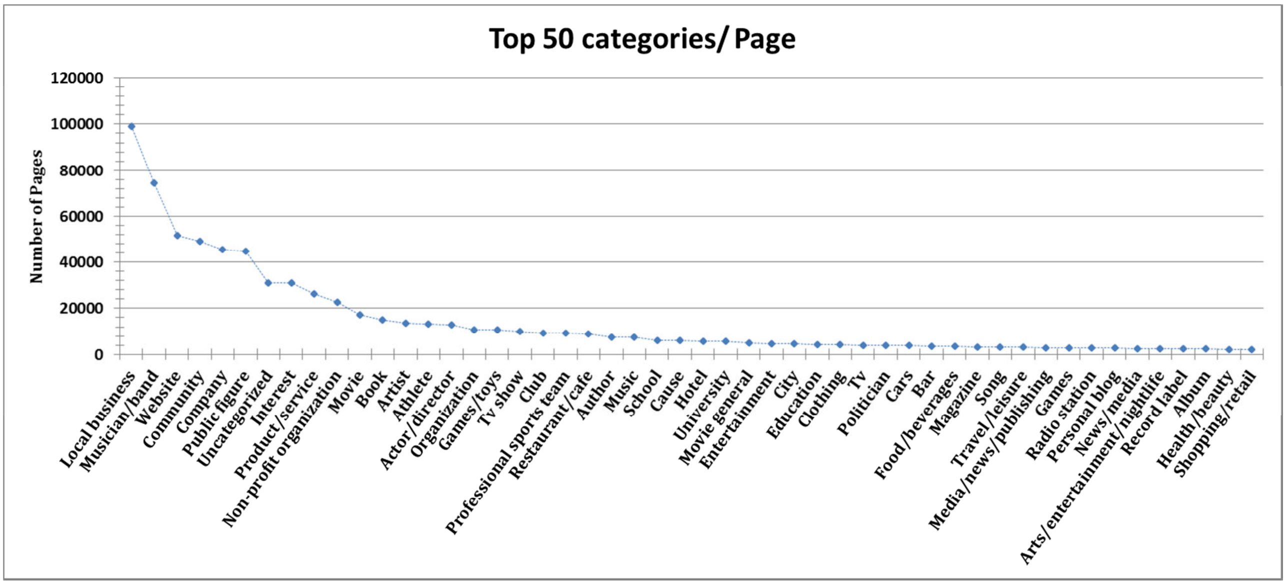


Figure 7: Top 50 Categories – Pages per Category

We observe that the curve on Figure 7 is giving us hints of a Power Law Distribution. The majority of categories have a small number of pages, while a couple of them have a large number of pages. We will examine the above theory by implementing the methods for *Power-law Distributions in Empirical Data* written by Aaron Clauset, Cosma R. Shalizi and M.E.J. Newman[17] and using the matlab code shared on Santa Fe Institute's site.[10]

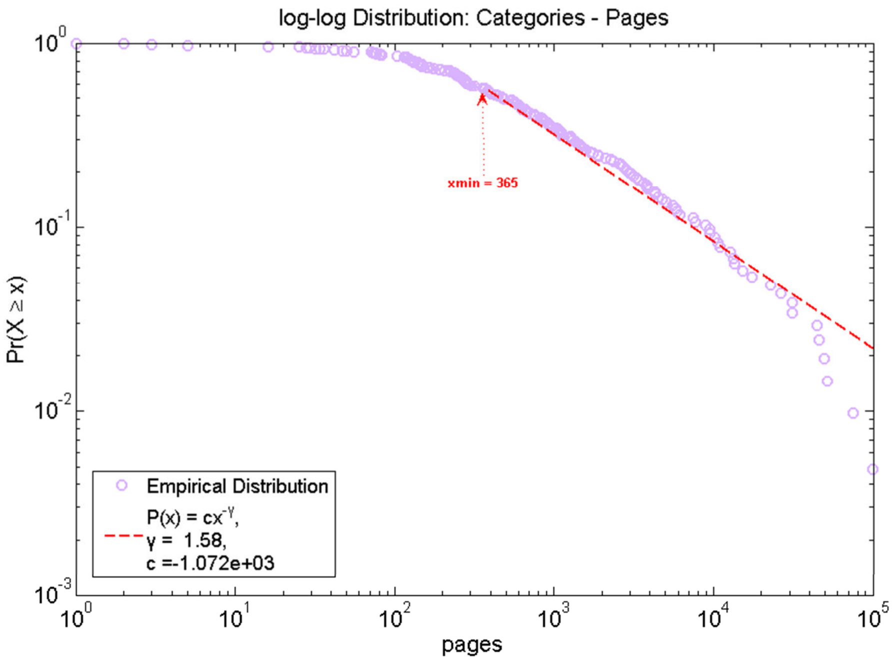


Figure 8: Top 50 Categories – Log –log Distribution: Categories Pages

---

[10] http://tuvalu.santafe.edu/~aaronc/powerlaws/

The data points' distribution of Categories – Pages share some power law characteristics.

2.2.2 Categories Popularity – Likes per Category

In the following table we present the likes distribution amongst the 50 most popular categories and the gender distribution. In Figure 9 we present a graph with the combined information.

#	Category	Likes	% Female	% Male	% Non-Disclosed
1	Local business	911801	58,40	39,08	2,52
2	Musician/band	588777	40,10	56,95	2,95
3	Community	404153	54,25	43,24	2,50
4	Public figure	388039	45,49	51,99	2,52
5	Website	268761	60,14	37,09	2,77
6	Company	239462	48,86	48,39	2,75
7	Tv show	200610	44,95	52,94	2,11
8	Movie	182276	43,62	54,14	2,24
9	Non-profit organization	168829	50,46	46,71	2,83
10	Product/service	159343	45,02	52,30	2,68
11	Actor/director	156420	43,35	54,13	2,52
12	Uncategorized	148464	45,73	51,74	2,53
13	Athlete	116839	21,98	75,73	2,30
14	Professional sports team	106246	23,38	74,15	2,47
15	Club	104779	44,84	52,61	2,55
16	Interest	101572	51,90	45,90	2,20
17	News/media	73475	42,55	54,65	2,80
18	Games	67671	47,53	50,43	2,04
19	Artist	64574	45,95	50,30	3,75
20	Games/toys	63360	31,01	67,00	1,99
21	Organization	60390	48,75	48,63	2,62
22	Entertainment	56403	49,52	48,10	2,38
23	Food/beverages	51975	47,32	50,44	2,24
24	Cause	47224	57,16	40,63	2,21
25	Restaurant/café	44087	43,65	53,15	3,20
26	Clothing	43408	61,28	35,60	3,13
27	Author	41497	45,23	51,90	2,87
28	Song	41365	60,31	37,29	2,41
29	Politician	40890	30,56	66,49	2,95
30	Radio station	39391	39,48	57,10	3,42
31	Book	38590	45,77	52,09	2,14
32	Personal blog	29232	51,80	45,56	2,65
33	Magazine	27253	44,52	52,04	3,44
34	Travel/leisure	26280	48,71	48,06	3,23
35	Hotel	25587	43,96	51,51	4,53
36	Education	24573	49,64	47,34	3,02
37	Bar	24226	41,72	55,30	2,99
38	Cars	23948	21,90	75,26	2,84
39	Fictional character	21313	49,81	47,98	2,21
40	Media/news/publishing	19264	42,51	54,57	2,92
41	Sports league	18763	22,42	75,28	2,30
42	Comedian	17517	42,83	54,79	2,37
43	Society/culture	15739	49,78	47,39	2,83
44	Musical genre	15713	30,72	67,64	1,64

45	Arts/entertainment/nightlife	15582	48,22	48,50	3,29
46	University	14567	46,28	51,05	2,67
47	Local/travel	14095	45,43	51,62	2,95
48	Sports venue	14084	26,16	71,26	2,58
49	City	13657	39,25	58,43	2,31
50	Museum/art gallery	13422	50,70	45,63	3,67

Table 4: Top 50 Categories – Likes per Category

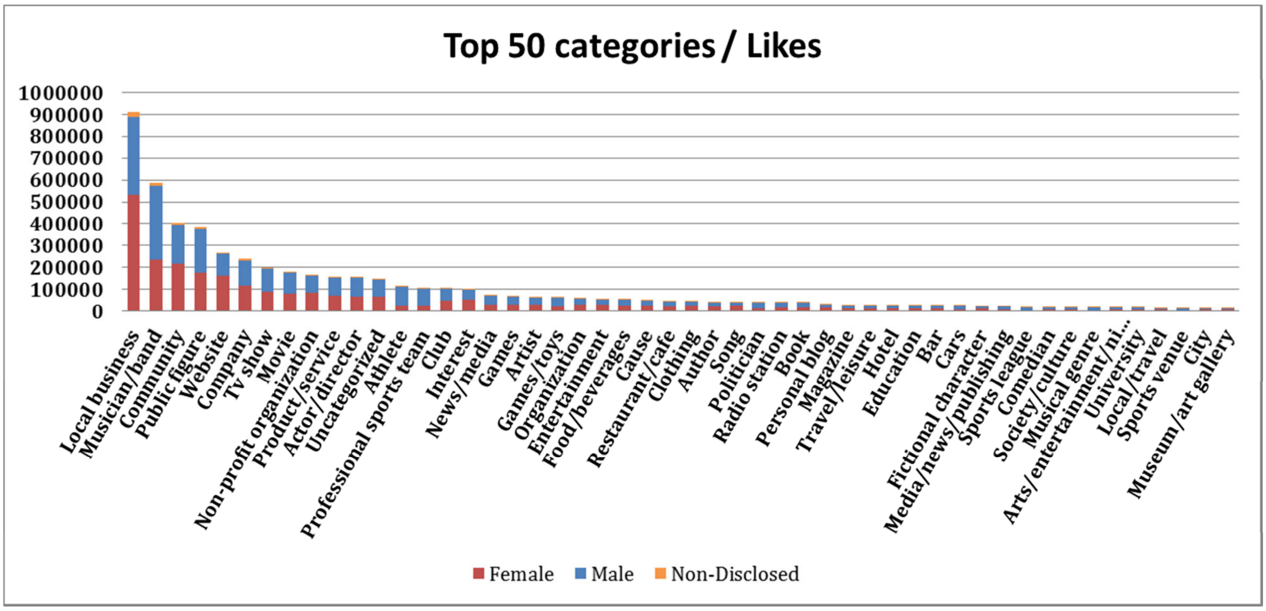


Figure 9: Top 50 Categories – Likes per Category

We compare the distribution of Categories – Likes with a Power Law Distribution. We can observe that part of the distribution follows the Power-Law.

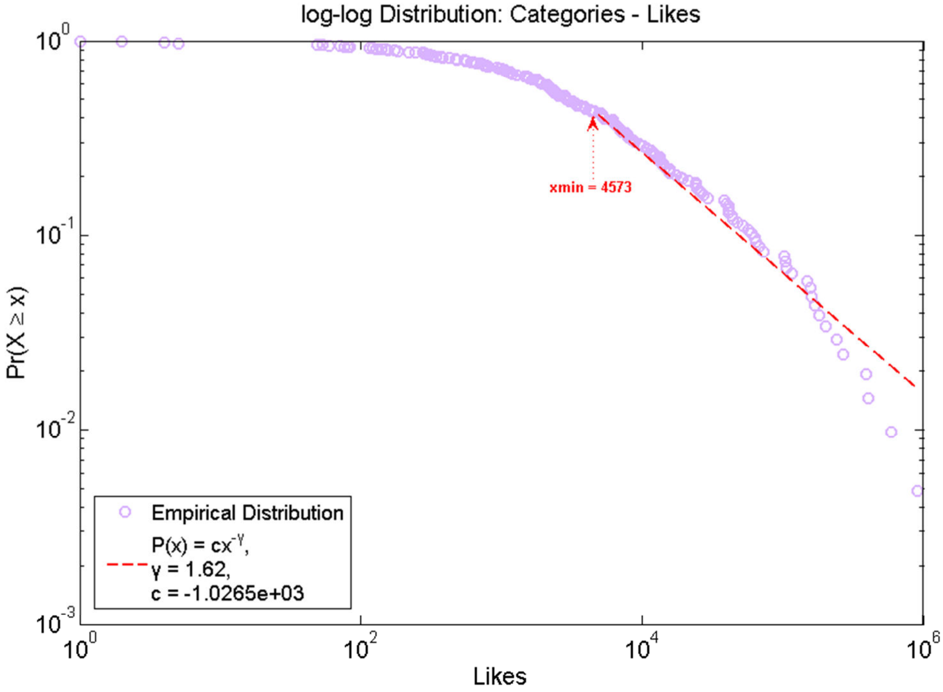


Figure 10: log-log Distribution: Categories - Likes

2.2.3 Top 50 Pages

Below we present the top 50 most liked pages and the breakdown according to gender. It is interesting to observe the difference of Male and Female preferences according to the category and content of the pages.

	Page Name	Category	Likes	Male %	Female %
1	Texas Hold em Poker	Games/toys	9351	74,93	22,29
2	ARKAS	Public figure	9194	56,80	40,22
3	Radio Arvila	Tv show	7830	56,67	40,88
4	Bring Them Back	Non-profit organization	7217	53,87	43,48
5	DES POIOS VLEPEI FOTOGRAFIES, PROFIL KAI POIOS SE EXEI DIAGRAPSEI !!(100%)	Product/service	7077	52,66	44,20
6	TO NHSI	Tv show	6869	31,68	65,92
7	VOITHEIA STIN AITI ARKEI MONO NA GINETAI MELOS (STEILE TO PANTOY)	Public figure	6604	47,85	49,14
8	H APANTISH MAS STIN TOYRKIKI SELIDA POY EXEI 200.000 MELH..EMEIS 300.000!!!	Community	6584	64,66	32,91
9	Acropolis Museum	Museum/art gallery	5851	49,48	47,65
10	Tzimis Panousis	Public figure	5830	73,09	23,60
11	Panathinaikos	Professional sports team	5726	66,78	30,41
12	EYTYXISMENOI MAZI	Tv show	5714	52,26	44,63
13	OOOOOOOOOOOOOOO......SOULTAN!!!!!!! (H EPISTROFH)!!!	Public figure	5576	59,40	38,40
14	THA TRELATHW .... THA PIDIXTW AP TO PARATHYRO !!!!!!!!!!	Tv show	5122	47,64	50,41
15	ELLINIKES DRAXMES	Non-profit organization	5115	63,07	33,98
16	Thanasis Veggos	Comedian	4807	71,29	25,77
17	Eva Mendes	Actor/director	4806	82,54	15,27
18	Ta KalyteraSymbainoynEkeiPoy Den to Perimeneis	News/media	4716	39,69	57,89
19	FRIENDS (TV Show)	Tv show	4628	46,87	51,40
20	LefteriaStaPaidiaTou Gamato.info	Public figure	4531	65,31	31,69
21	DIADWSE TO: Oloi se miaselidagia tin mnimitou Alexi Grigoropoulou	Public figure	4449	50,37	47,13
22	Giannis Mpezos	Actor/director	4311	66,39	30,62
23	AS VOITHISOUME STIN PRAXI TA ADESPOTA (STEILE TO PANTOY)	Non-profit organization	4308	42,20	54,43
24	FarmVille	Games	4263	45,06	52,85
25	Monica Bellucci	Actor/director	4243	76,76	20,36
26	THELW NA TAXIDEPSW SE OLO TON KOSMO!!	Travel/leisure	4229	36,13	61,22
27	Michael Jackson	Musician/band	4172	56,54	41,35
28	Dimitris Mitropanos	Public figure	4083	58,56	38,94
29	Iron Mike Zambidis	Athlete	4044	75,82	21,66
30	Dr. House	Fictional character	4021	52,77	45,41
31	OXI STOYS ROYFIANOYS	Public figure	3962	59,97	37,38
32	Eimaistonkosmomougiati den mouaresei o dikossas....!!!	Local business	3909	37,78	59,68
33	Filipidis Petros Funs	Actor/director	3865	59,04	38,11
34	ThunderCats	Tv show	3848	68,87	28,35
35	Maria Solomou	Actor/director	3838	61,72	35,64
36	Eleni Rantou	Actor/director	3805	42,08	55,69
37	Megan Fox	Actor/director	3750	78,93	19,17

38	STAMATISTE PIA NA DILITIRIAZETE TA ZWAKIA	Local business	3739	39,66	57,66
39	Mixalis Xatzigiannis	Musician/band	3735	32,34	64,31
40	OLYMPIAKOS. gia mia zwi	Professional sports team	3731	66,68	30,42
41	50 - 50	Tv show	3648	55,07	42,60
42	Lacta	Food/beverages	3603	34,39	63,36
43	Sakis Rouvas	Public figure	3546	21,66	75,55
44	Thafygw RE... tha paw allou!!!!	Personal blog	3533	44,98	52,59
45	Bob Marley	Musician/band	3518	62,56	34,51
46	MARKOS SEFERLIS	Local business	3507	69,58	28,31
47	Pote min odigateenwexetepiei....DiavastemiaSYGKLONISTIKI istoria!	Local business	3502	49,26	48,20
48	Lelos (Ela Liza! Mpanana!)	Public figure	3498	53,00	44,77
49	South Park	Tv show	3465	67,88	30,13
50	MERA XWRIS XAMOGELO EINAI XAMENI MERA	Personal blog	3393	36,46	61,10

Table 2.2.3: Top 50 Pages

We provide a bar graph, with the combined information in Figure

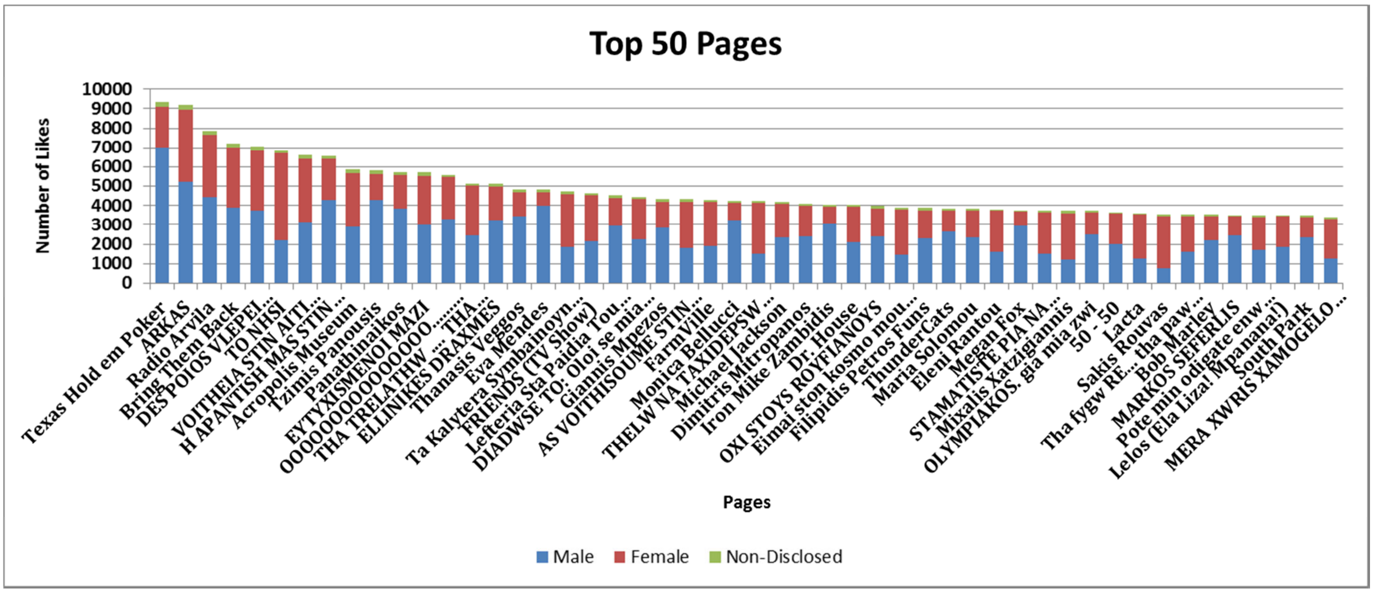


Figure 11: Top 50 Pages

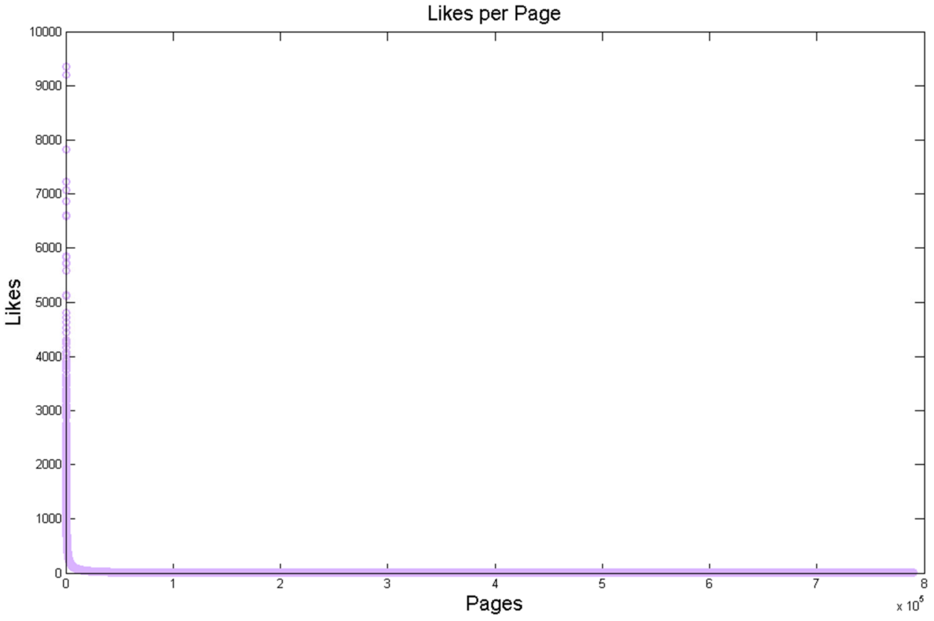


Figure 12: All Pages and their Like count in Graph

In Figure 13 we show that the Likes per Pages log-log Distribution aligns with the Power Law distribution for a part of it.

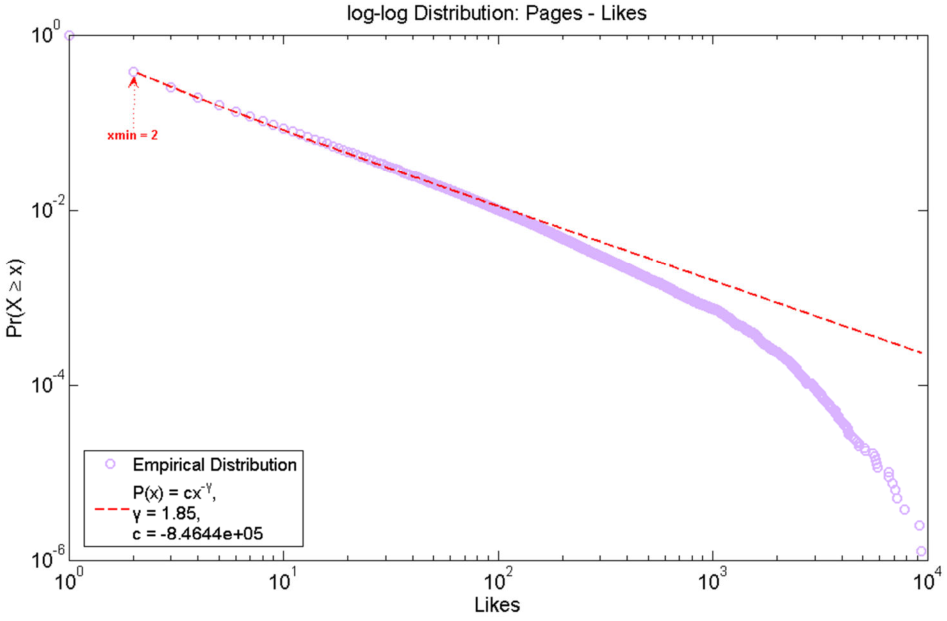


Figure 13: log-log Distribution: Pages - Likes

2.2.4 TOP 10 Pages Men

We provide a Top 10 matrix with the most liked pages by male users. We can observe the similarity with the general Top 50 matrix and also the rise of specific pages that appeal mostly to the male population.

#	**Page Name**	**Category**	**Likes**	**Men**	**Women**
1	Texas Hold em Poker	Games/toys	9351	7007	2084
2	ARKAS	Public figure	9194	5222	3698
3	Radio Arvila	Tv show	7830	4437	3201
4	Tzimis Panousis	Public figure	5830	4261	1376
5	H APANTISH MAS STIN TOYRKIKI SELIDA POY EXEI 200.000 MELH..EMEIS 300.000!!!	Community	6584	4257	2167
6	Eva Mendes	Actor/director	4806	3967	734
7	Bring Them Back	Non-profit organization	7217	3888	3138
8	Panathinaikos	Professional sports team	5726	3824	1741
9	DES POIOS VLEPEI FOTOGRAFIES, PROFIL KAI POIOS SE EXEI DIAGRAPSEI !!(100%)	Product/service	7077	3727	3128
10	Thanasis Veggos	Comedian	4807	3427	1239

Table 5: Top 10 Men

2.2.5 TOP 10 Pages Women

We provide a Top 10 matrix with the most liked pages by female users. Again we can observe the similarity with the general Top 50 matrix and also the rise of specific pages that appeal mostly to female users.

#	Page Name	Category	Likes	Men	Women
1	TO NHSI	Tv show	6869	2176	4528
2	ARKAS	Public figure	9194	5222	3698
3	VOITHEIA STIN AITH ARKEI MONO NA GINETAI MELOS (STEILE TO PANTOY)	Public figure	6604	3160	3245
4	Radio Arvila	Tv show	7830	4437	3201
5	Bring Them Back	Non-profit organization	7217	3888	3138
6	DES POIOS VLEPEI FOTOGRAFIES, PROFIL KAI POIOS SE EXEI DIAGRAPSEI !!(100%)	Product/service	7077	3727	3128
7	Acropolis Museum	Museum/art gallery	5851	2895	2788
8	Ta KalyteraSymbainoynEkeiPoy Den to Perimeneis	News/media	4716	1872	2730
9	Sakis Rouvas	Public figure	3546	768	2679
10	THELW NA TAXIDEPSW SE OLO TON KOSMO!!	Travel/leisure	4229	1528	2589

Table 6: Top 10 Women

2.3 Dataset II (Uniform Evolution Sample) Analysis

We provide 4 matrixes that demonstrate the evolution in simple statistics of the graph during the 10 day period from 18/05/11 to 27/05/11. We intend to examine the diffusion of likes and friendships in direct comparison with Dataset III.

2.3.1 Nodes - Edges Evolution

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Level 0	500	500	500	500	500	500	500	500	500	500
Level 1	187151	187176	187286	186244	185756	186640	186750	186861	186933	186375
Total Nodes	187651	187676	187786	186744	186256	187140	187250	187361	187433	186875
Difference		25	110	-1042	-488	884	110	111	72	-558
Lost		-112	-191	-1345	-860	-177	-260	-184	-201	-815
New		+137	+301	+303	+372	+1061	+370	+295	+273	+257
%		0,01%	0,06%	-0,55%	-0,26%	0,47%	0,06%	0,06%	0,04%	-0,30%

Table 7: Dataset II, Nodes Evolution

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Total Edges	208756	208796	208947	207765	207279	208231	208386	208531	208629	207838
Difference		40	151	-1182	-486	952	155	145	98	-791
Lost		-131	-215	-1526	-911	-197	-288	-214	-220	-1096
New		+171	+366	+344	+425	+1149	+443	+359	+318	+305
%		0,02%	0,07%	-0,57%	-0,23%	0,46%	0,07%	0,07%	0,05%	-0,38%

Table 8: Dataset II, Edges Evolution

2.3.2 Pages – Likes Evolution

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Explored Nodes	334	334	334	331	331	331	331	332	332	331
# Unique Pages	21104	21164	21133	20156	20123	20193	20213	20220	20423	20414
Difference		60	-31	-977	-33	70	20	7	203	-9
Unliked		-15	-55	-1055	-52	-11	-25	-265	-107	-138
Liked		+75	+24	+78	+19	+81	+45	+272	+310	+129
%		0,28%	-0,15%	-4,62%	-0,16%	0,35%	0,1%	0,03%	1,00%	-0,04%

Table 9: Dataset II - Pages Evolution

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Explored Nodes	334	334	334	331	331	331	331	332	332	331
Likes	3099	33195	33138	31521	31470	31564	31610	31599	31926	31802
Difference		96	-57	-1617	-51	94	46	-11	327	-124
New Likes		+131	+36	+147	+32	+113	+74	+438	+519	+231
Removed Likes		-35	-93	-1764	-83	-19	-28	-449	-192	-355
%		0,29%	-0,17%	-4,88%	-0,16%	0,30%	0,15%	-0,03%	1,03%	-0,39%

Table 10: Dataset II - Likes Evolution

2.4 Dataset III (Friends Evolution Sample) Analysis

In the following sections, statistics and graphs are provided regarding the evolution of Dataset III graph, during the 20 day period from 08/05/11 to 27/05/11.

2.4.1 Nodes - Edges Evolution

We observe the daily change of nodes population and their respective edges. On a daily basis we crawl our source node on level 0 and collect its friends that reside on level 1. We then repeat this procedure collecting the "friends of friends" of our source node that reside on level 2. The sum of the source node, its friends and friends of friends equals the Total Nodes of the graph. We count the added and removed nodes from one day to the next and the difference in the Total Population. We can observe that the biggest alterations result when a node is removed or added on the first level (friends) as its friends list is added or removed consequently. We observe that the percentage of the difference of the nodes and the edges is almost equal as most of the nodes that are added or removed are "leaves" (nodes with one edge).

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11
Level 0	1	1	1	1	1	1	1	1	1	1	1
Level 1	344	343	343	343	342	342	342	342	341	342	342
Level 2	100045	99745	99785	99709	99965	99996	100030	100118	100092	100388	100253
Total Nodes	100390	100089	100129	100053	100308	100339	100373	100461	100434	100731	100596
Difference		-301	40	-76	255	30	34	88	-27	297	-135
Removed		-391	-102	-223	-223	-61	-56	-87	-143	-80	-360
Added		+90	+142	+147	+478	+92	+90	+175	+116	+377	+225
%		-0,30%	0,04%	-0,08%	0,25%	0,03%	0,03%	0,09%	-0,03%	0,30%	-0,13%

	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20
Level 0	1	1	1	1	1	1	1	1	1	1	1
Level 1	342	342	342	343	343	343	343	342	343	343	341
Level 2	100388	100253	100265	100686	100748	100804	100870	100927	100981	101300	100715
Total Nodes	100731	100596	100608	101030	101092	101148	101214	101270	101325	101644	101057
Difference		-135	12	422	62	56	66	56	55	319	-587
Removed		-360	-49	-144	-61	-90	-83	-118	-76	-57	-707
Added		+225	+61	+566	+123	+146	+149	+174	+131	+376	+120
%		-0,13%	0,01%	0,42%	0,06%	0,06%	0,07%	0,06%	0,05%	0,31%	-0,58%

Table 11 – Dataset III - Nodes Evolution

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11
Total Edges	124267	123801	123865	123697	124113	124158	124197	124311	124223	124564	124403
Difference		-466	64	-168	416	45	39	114	-88	341	-161
Lost		-580	-115	-335	-268	-77	-67	-100	-233	-93	-431
New		+114	+179	+167	+684	+122	+106	+214	+145	+434	+270
%		-0,37%	0,05%	-0,14%	0,34%	0,04%	0,03%	0,09%	-0,07%	0,27%	-0,13%

	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20
Total Edges	124564	124403	124419	124907	124973	125040	125121	125178	125231	125618	124749
Difference		-161	16	488	66	67	81	57	53	387	-869
Lost		-431	-56	-155	-74	-104	-91	-146	-105	-65	-1011
New		+270	+72	+643	+140	+171	+172	+203	+158	+452	+142
%		-0,13%	0,01%	0,39%	0,05%	0,05%	0,06%	0,05%	0,04%	0,31%	-0,69%

Table 12: Dataset III - Edges Evolution

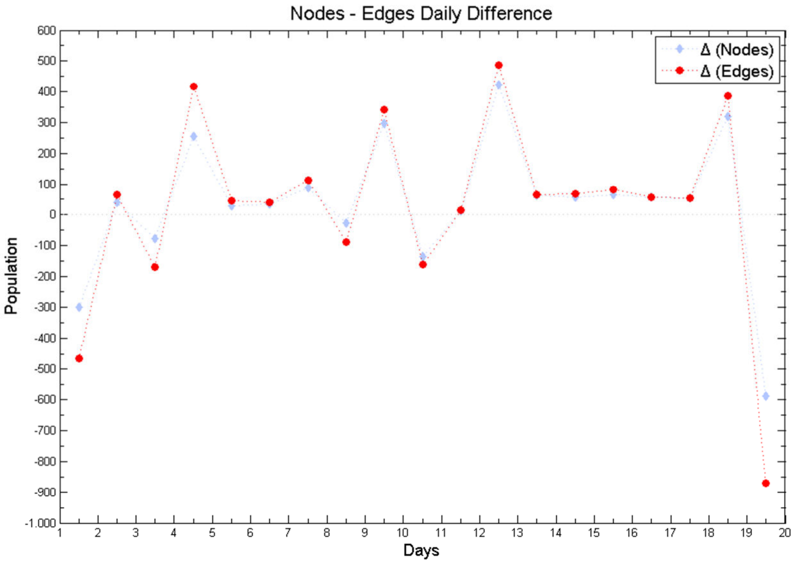


Figure 14: Nodes, Edges Daily Difference

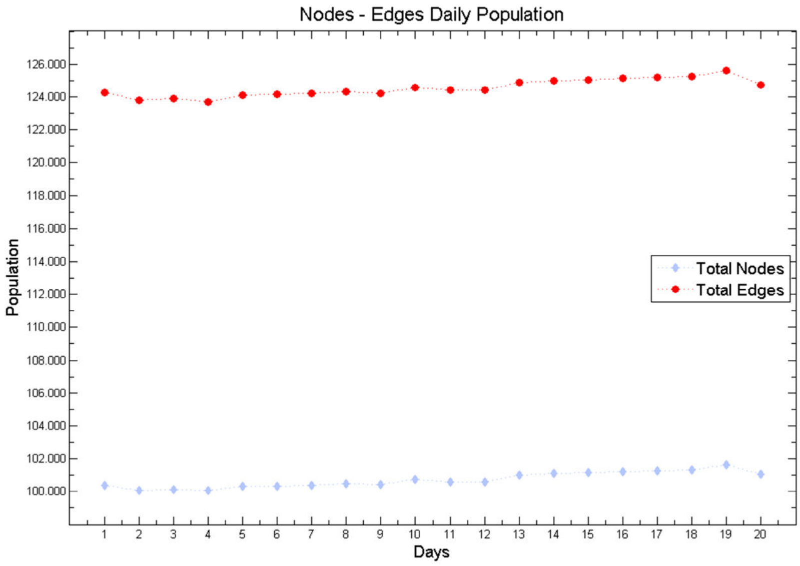


Figure 15: Nodes, Edges Population Evolution

2.4.2 Pages - Likes Evolution

On a daily basis we gather the “Likes” of the 1st level nodes (Friends). The FB API is not disclosing the likes of all the users, while a portion of users do not share them intentionally. The number of explored nodes and Pages they liked is represented on the Tables below.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11
Explored Nodes	280	280	280	278	276	277	276	276	276	276	275
# Unique Pages	13550	13349	13361	13373	13346	13386	13360	13385	13376	13350	13333
Difference		-1	12	12	-27	40	-26	25	-9	-26	-17
Unliked		-9	-13	-17	-58	-8	-45	-24	-60	-40	-50
Liked		+8	+25	+29	+31	+48	+19	+49	+51	+14	+33
%		-1,48%	0,09%	0,09%	-0,20%	0,30%	-0,19%	0,19%	-0,07%	-0,19%	-0,13%

	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20
Explored Nodes	276	275	277	278	278	278	278	277	278	278	278
# Unique Pages	13350	13333	13375	13438	13453	13467	13476	13484	13326	13509	13529
Difference		-17	42	63	15	14	9	8	-158	183	20
Unliked		-50	-4	-20	-8	-8	-8	-18	-173	-12	-7
Liked		+33	+46	+83	+23	+22	+17	+26	+15	+195	+27
%		-0,13%	0,32%	0,47%	0,11%	0,10%	0,07%	0,06%	-1,17%	1,37%	0,15%

Table 13: Dataset III - Pages Evolution

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11
Explored Nodes	280	280	280	278	276	277	276	276	276	276	275
Likes	20143	20143	20159	20178	20123	20190	20140	20170	20156	20118	20068
Difference		0	16	19	-55	67	-50	30	-14	-38	-50
New Likes		+12	+34	+41	+44	+76	+26	+88	+100	+32	+57
Rem. Likes		-12	-18	-22	-99	-9	-76	-58	-114	-70	-107
%		0,00%	0,08%	0,09%	-0,27%	0,33%	-,25%	,15%	-0,07%	-,19%	-0,25%

	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20
Explored Nodes	276	275	277	278	278	278	278	277	278	278	278
Likes	20118	20068	20161	20291	20327	20355	20373	20380	20092	20432	20461
Difference		-50	93	130	36	28	18	7	-288	340	29
New Likes		+57	+106	+156	+48	+36	+27	+37	+37	+358	+47
Rem. Likes		-107	-13	-26	-12	-8	-9	-30	-325	-18	-18
%		-0,25%	0,46%	0,64%	0,18%	0,14%	0,09%	0,03%	-1,41%	1,69%	0,14%

Table 14: Dataset III - Likes Evolution

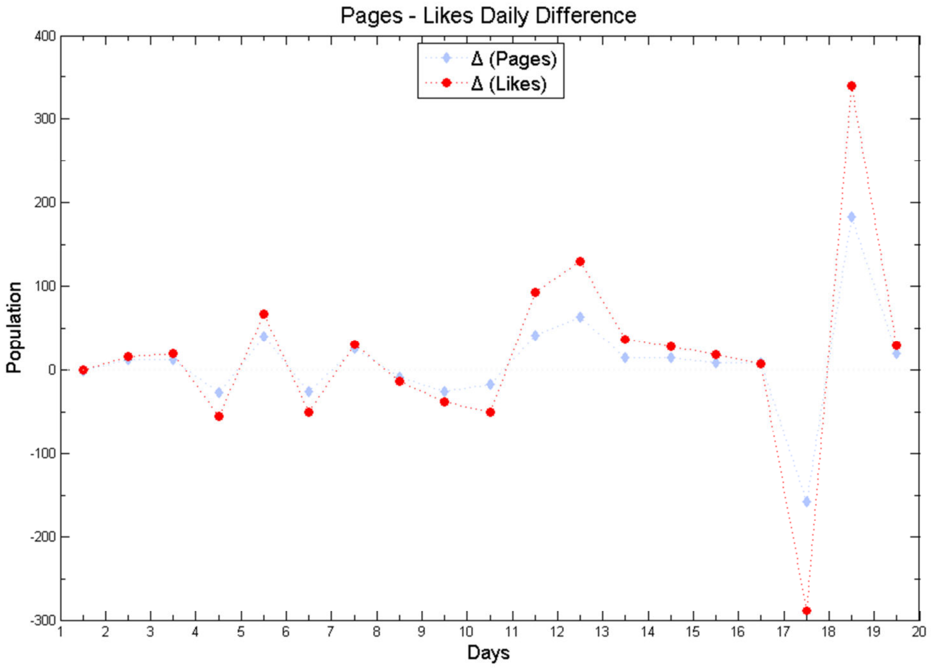


Figure 16: Pages, Likes Daily Difference

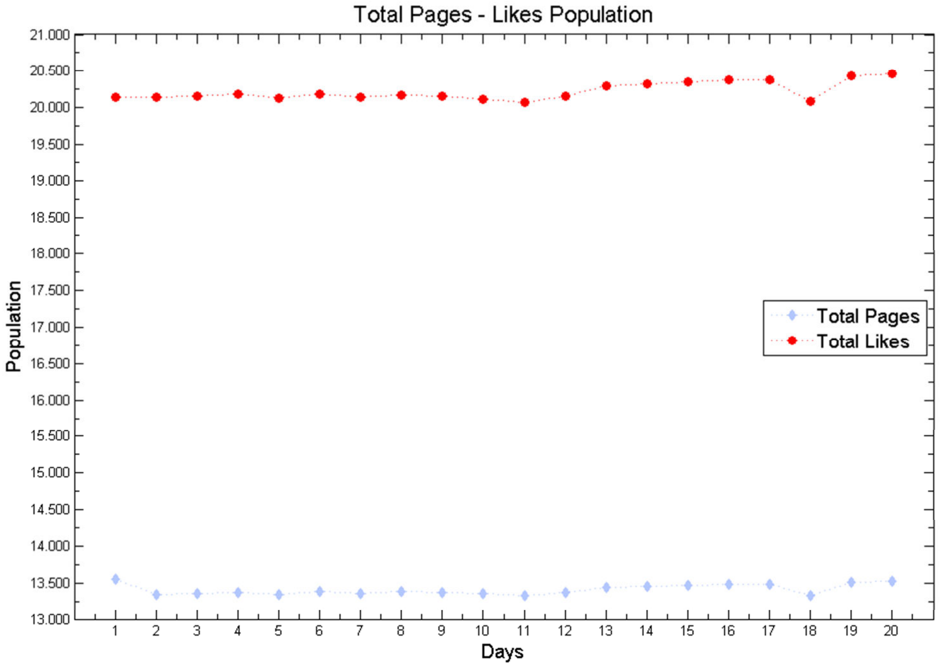


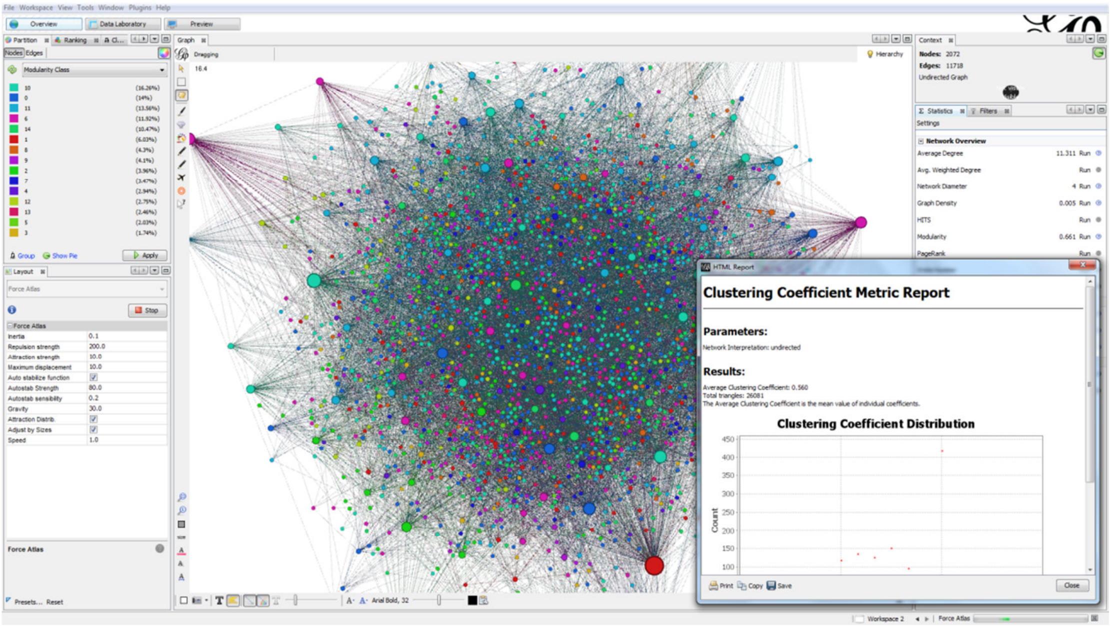
Figure 17: Pages, Likes Population Evolution

3 Graph Analysis

In an aim to understand and explore further the collected data we use Gephi for our Graph Analysis purposes.

3.1 Graph Analysis Software

Gephi is an open-source software for visualizing and analyzing large network graphs. Gephi uses a 3D render engine (OpenGL) to display graphs in real-time, allowing to easily explore, analyse, spatialise, filter, clusterize, manipulate all types of graphs. Gephi's fast graph visualization engine enables users to understand and discover patterns in large graphs, while the platform can handle networks up to 50K nodes and 500K edges, can iterate through visualization using dynamic filtering and provide rich tools for meaningful graph manipulation. On top of it, Gephi provides state-of-the-art layout algorithms and most common metrics for social network analysis (SNA) and scale-free networks such as: Betweenness, Closeness, Diameter, Clustering Coefficient, Average shortest path, PageRank, HITS, Community detection (Modularity), Random generators, etc.



Modularity Class

■	10	(16.26%)
■	0	(14%)
■	11	(13.56%)
■	6	(11.92%)
■	14	(10.47%)
■	1	(6.03%)
■	8	(4.3%)
■	9	(4.1%)
■	2	(3.96%)
■	7	(3.47%)
■	4	(2.94%)
■	12	(2.75%)
■	13	(2.46%)
■	5	(2.03%)
■	3	(1.74%)

Force Atlas

Inertia	0.1
Repulsion strength	200.0
Attraction strength	10.0
Maximum displacement	10.0
Auto stabilize function	☑
Autostab Strength	80.0
Autostab sensibility	0.2
Gravity	30.0
Attraction Distrib.	☑
Adjust by Sizes	☑
Speed	1.0

Nodes: 2072
Edges: 11718
Undirected Graph

Network Overview

Average Degree	11.311	Run
Avg. Weighted Degree		Run
Network Diameter	4	Run
Graph Density	0.005	Run
HITS		Run
Modularity	0.661	Run
PageRank		Run

HTML Report

Clustering Coefficient Metric Report

Parameters:

Network Interpretation: undirected

Results:

Average Clustering Coefficient: 0.560
Total triangles: 26081
The Average Clustering Coefficient is the mean value of individual coefficients.

Clustering Coefficient Distribution

Figure 18: A screenshot of Gephi version 0.8 Alpha

3.2 Analyzing Dataset I

We the help of Gephi we will analyze the Facebook Graph taken from Dataset I. For our Graph we will provide drawings and measure properties of the graph such as clustering-coefficient, average degree-distribution, Betweenness Centrality Distribution, Closeness Centrality Distribution, Eigenvector-centralities and Modularity.

3.2.1 Visualization

We import our graph that consists of 124.267 edges and 100.390 nodes. We filter out all leaves (nodes with a single edge), resulting in a 13.790 nodes, 37.667 edges graph. Below we can see the visualization of part of the graph. Size of the node, denotes degree (Bigger node = Higher Degree), whilst color of the nodes denote the communities that were discovered (26 communities).

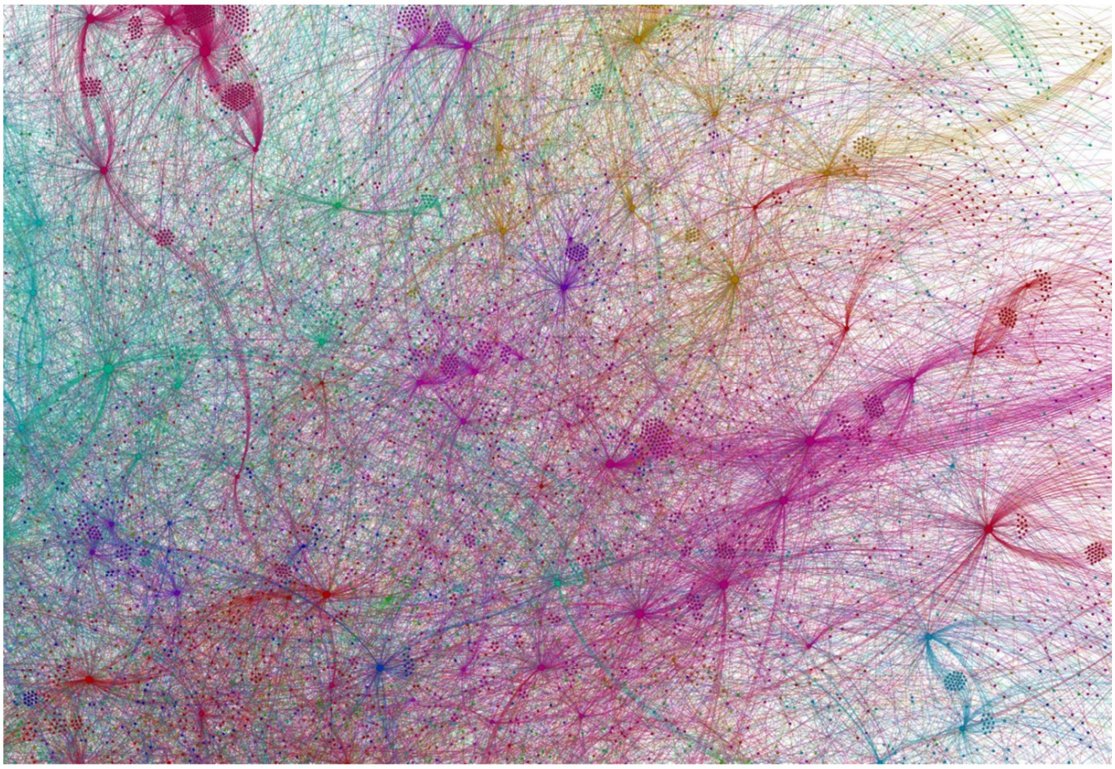


Figure 19: Part of Facebook's Graph

3.2.2 Metrics

We display some of the metrics we are able to produce with the help of Gephi.

3.2.2.1 Average Degree

We calculate and present the Degree Distribution of the 1st level nodes (Friends) of the graph.

- Average Degree: 366.093

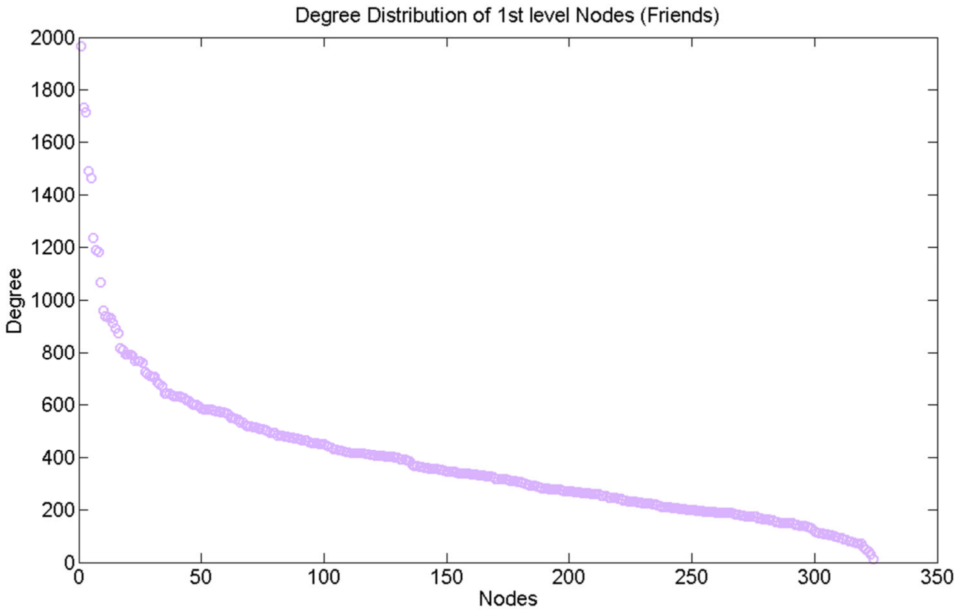


Figure 20: Degree Distribution of 1st level Nodes (Friends)

Nodes' degree in real-world, large scale social networks often follow a *power law distribution*. This could be no different in a OSN such as Facebook, so we will investigate this theory in Figure 21.

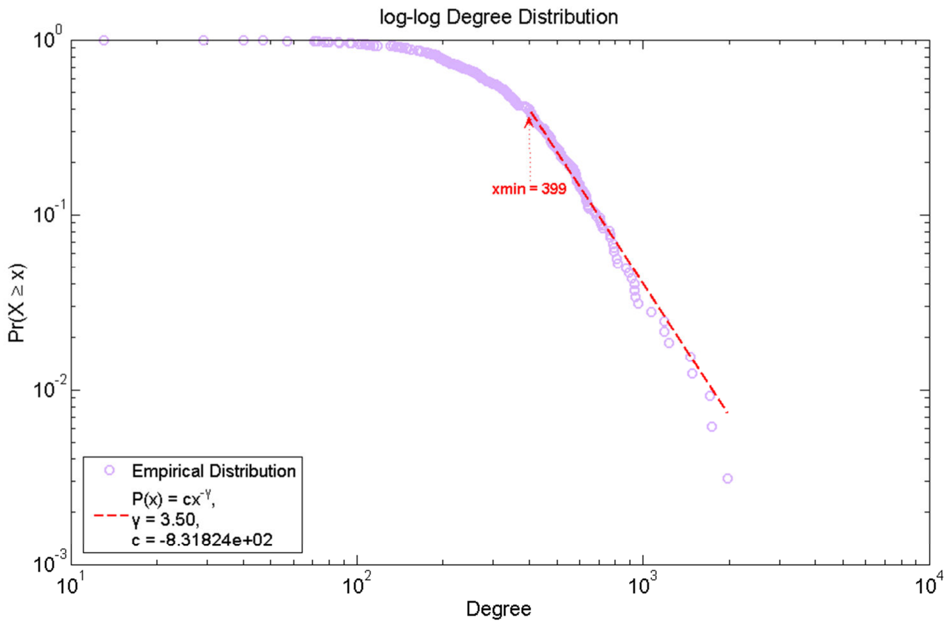


Figure 21: log-log Degree Distribution

We were expecting the Degree Distribution to be a Power Law Distribution. We observe to a great extent.

3.2.2.2 Average Path Length

The algorithm used here is based on Ulrik Brandes's, "A Faster Algorithm for Betweenness Centrality publication in Journal of Mathematical Sociology"[18].

Ulrik Brandes introduces more efficient algorithms based on a new accumulation technique that integrates well with traversal algorithms solving the single-source shortest-paths problem, and thus exploiting the sparsity of typical instances. This extends the range of networks for which betweenness centrality can be computed while being able to evaluate simultaneously all standard centrality indices based on shortest paths, thus reducing time and space requirements.

Centrality distributions regarding Closeness, Betweenness and Eigenevector are presented below.

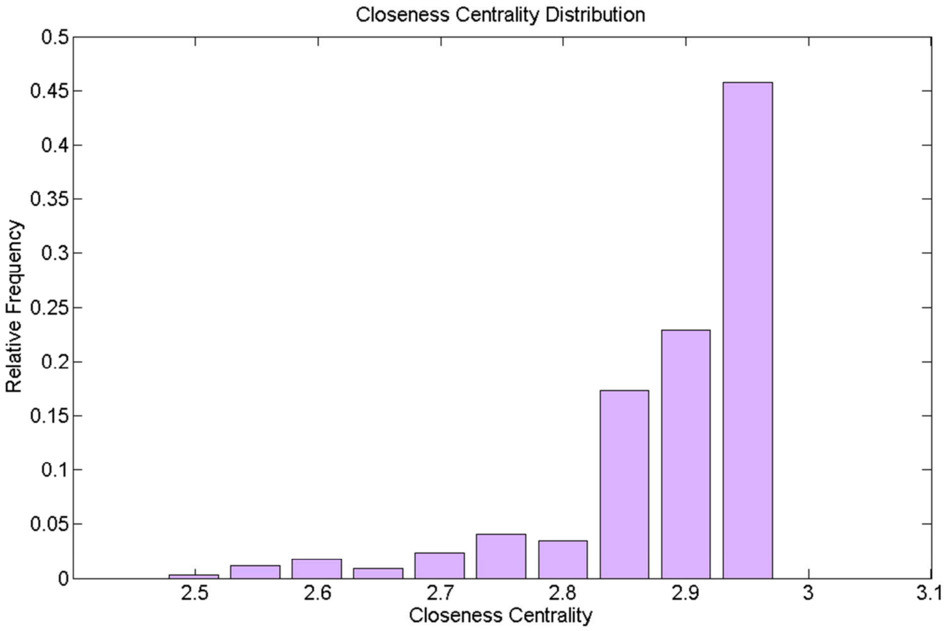


Figure 22: Closeness Centrality Distribution

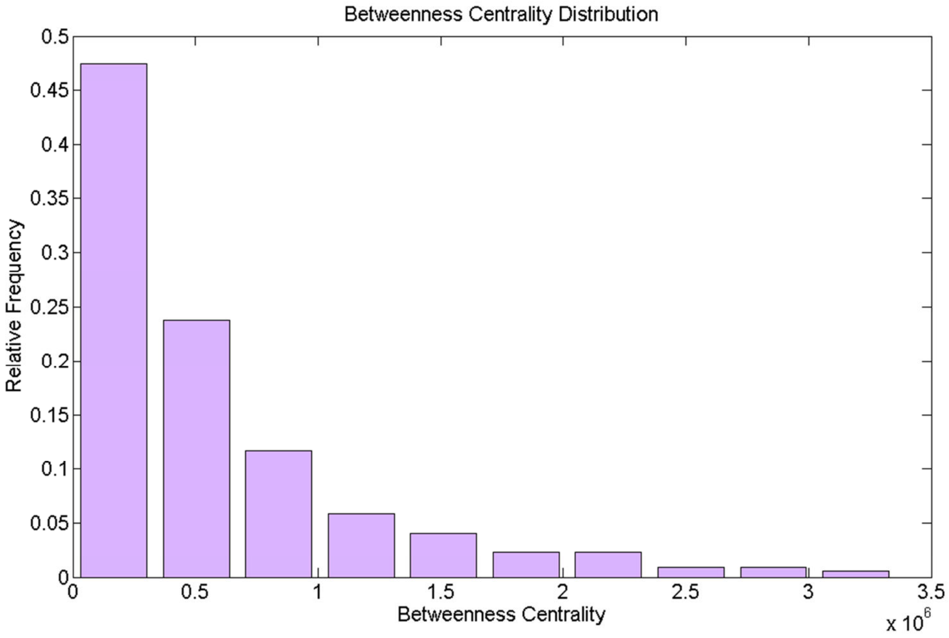


Figure 23: Betweenness Centrality Distribution

3.2.2.3 Eigenvector Centrality

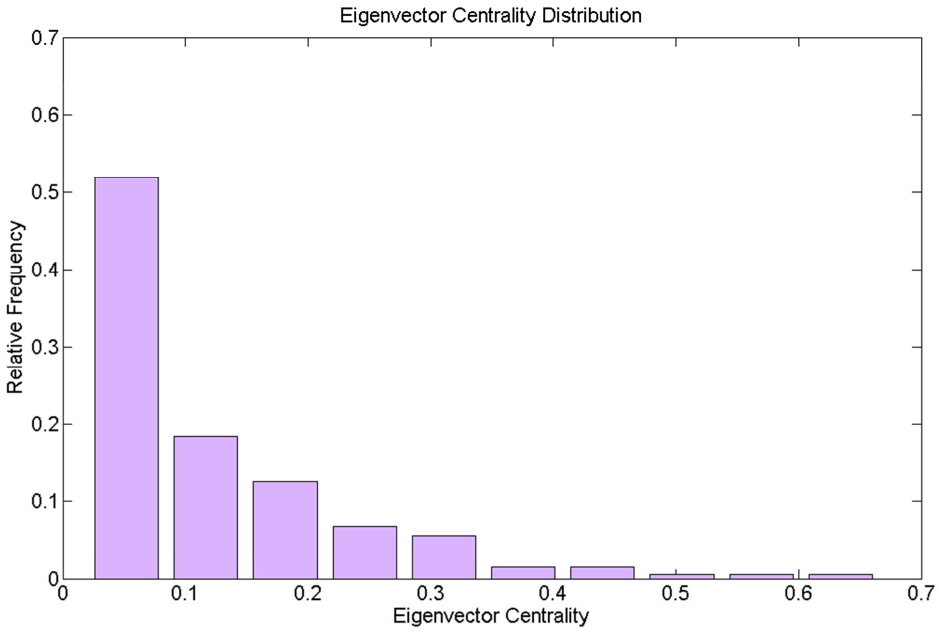


Figure 24: Eigenvector Centrality Distribution

3.2.2.4 Clustering Coefficient

Based on Matthieu Latapy's, Main-memory Triangle Computations for Very Large (Sparse (Power-Law)) Graphs, in Theoretical Computer Science (TCS) 407 (1-3), pages 458-473, 2008, we present the Clustering Coefficient of the 1st level nodes of our graph.

Average Clustering Coefficient: 0.0935 (The average Clustering Coefficient is the mean value of all individual coefficients.)

Total Triangles: 25751

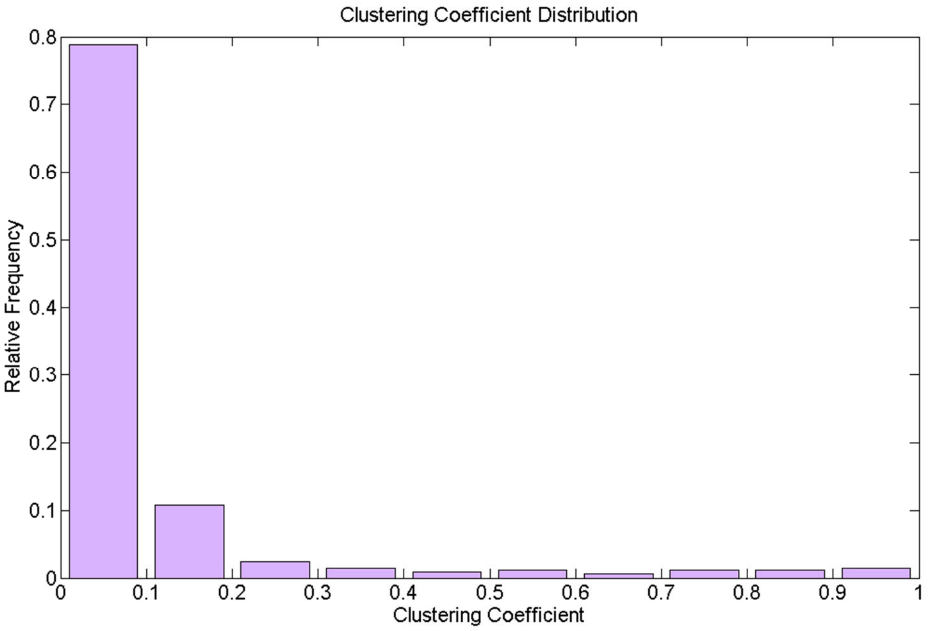


Figure 25: Clustering Coefficient Distribution

3.2.2.5 Modularity

The Modularity algorithm is based on the method published by Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre with the title “Fast unfolding of communities in large networks”. [19]

The Louvain Method analyzes networks of exceptional size very fast as analyzing a network of 2 million nodes takes approximately 2 minutes.

Two steps are repeated iteratively in the Louvain Method until a maximum of modularity is attained:

I. The Louvain Method looks for "small" communities by optimizing modularity in a local way.
II. Then it aggregates nodes of the same community, building a new network of communities.

Running the program results in several partitions, communities of small sizes. As the process iterates, larger and larger communities are found due to the aggregation mechanism, leading to the hierarchical decomposition of the network.

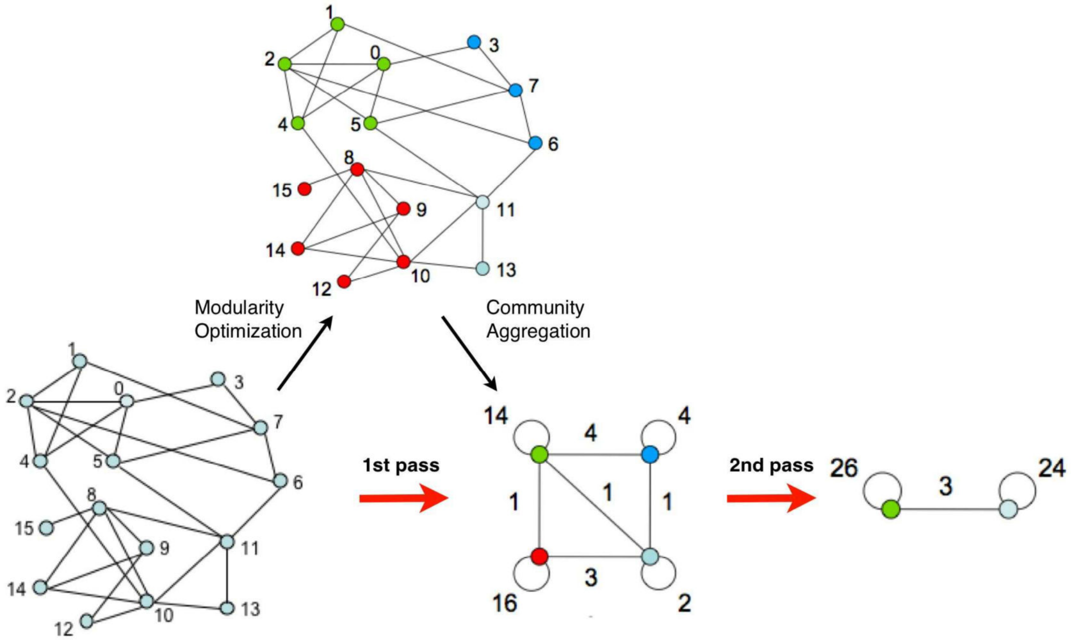


Figure 26: Visualization of the steps of the Louvain algorithm[11]

When implementing the algorithm on the full graph, 26 communities are unfolded with a modularity of 0.684.

In order to examine the effectiveness of the algorithm, we will import in Gephi only the first two crawled levels of nodes and edges (Source node and Friends).

With a modularity of 0.614, nine communities are returned. The graph visualization and the communities can be seen on figure 27.

---

[11] Visualization figure from Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks.

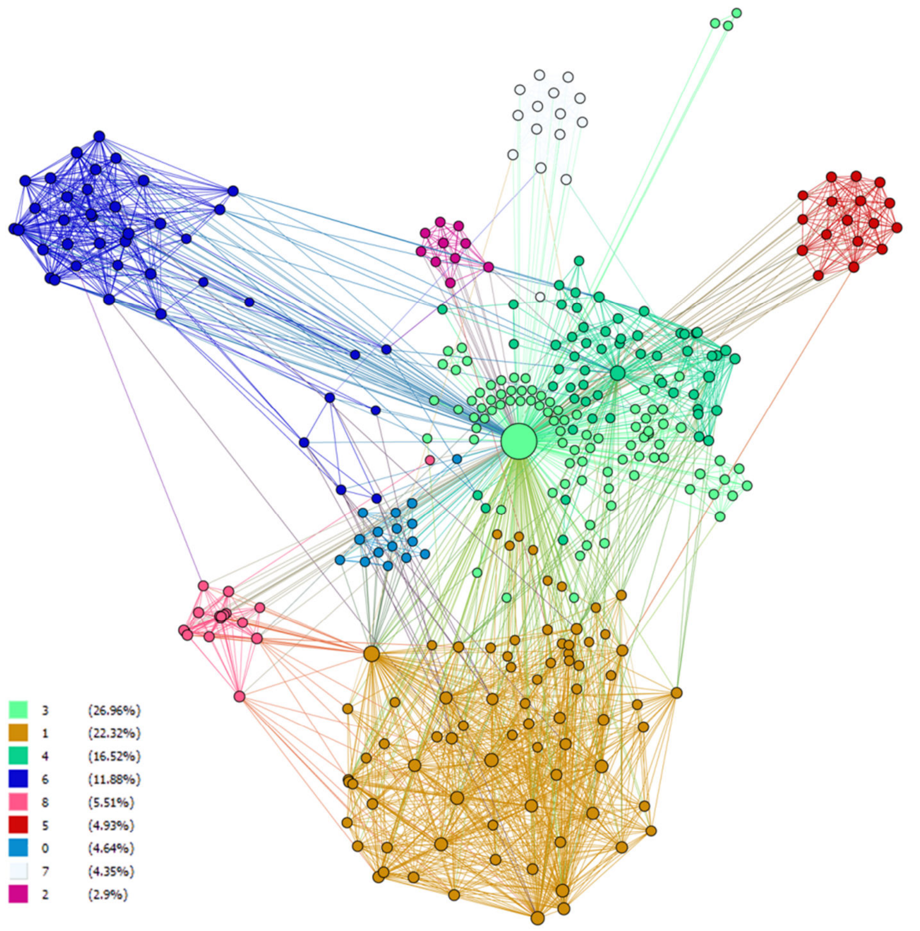


Figure 27: Friends Graph

It was discovered that the unfolded communities correspond to real life social groups of friends and acquaintances of the author. In particular:

🟪	0	Author's Friends from an Exchange Student Program
🟧	1	Authors group of close Friends 1
🟪	2	Author's Ex Coworkers
🟩	3	Author and random friends that could not be teamed together
🟩	4	Authors group of close Friends 2
🟥	5	Author's Friends from his Army Service
🟦	6	Author's Friends from his School
⬜	7	Author's Friends from his MSc Studies
🟥	8	Author's Friends from his Bachelor Studies

Table 15: Communities

The author observes all his co-workers grouped together under the same community, his old classmates constituting another community, his friends and acquaintances forming "cliques" that greatly depict real life associations and so on. These communities are determined by the density of links between members.

Nevertheless, are these communities sharing any other characteristics, other than dense connections between their members?

If we examine the “Ex co-workers” community, a page promoting their employee brand is quite popular among them in comparison to other communities. Likewise we noticed old Schoolmates liking a page that refers to their school and another promoting their district/ hometown. We became interested to identify if in all communities existed pages that highly identified and differentiated each community from another. If yes, could a community detection mechanism rely on the users’ common preferences (likes) and be more effective?

We aim to discover the above and present our findings in a future report.

4 Conclusion

Online Social Networks are without a doubt one of the most intriguing phenomena of the last years. For this study the OSN under analysis was Facebook, the most popular OSN with over 750 million users worldwide. In order to obtain a sample from Facebook we used Web Data Mining techniques and crawled the Facebook graph using the BFS technique. We applied SNA methods on the collected data and explored the graph of FB friendships. Future developments include studying the community discovery mechanisms and discovering the connection between users’ likes within a community structure and ways to improve community detection based on users’ likes.

REFERENCES

[1] Gemeinschaft und Gesellschaft, Leipzig: Fues's Verlag, 2nd ed. 1912, 8th edition, Leipzig: Buske, 1935 (reprint 2005, Darmstadt: Wissenschaftliche Buchgesellschaft)

[2] Durkheim, Emile. The Division of Labor in Society. Trans. Lewis A. Coser. New York: Free Press, 1997, pgs. 39, 60, 108.

[3] M. Newman. Power laws, Pareto distributions and Zipf's law. Contemporary physics, 46(5):323–352, 2005.

[4] Matthew O. Jackson, Social and Economic Networks, 2008

[5] Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

[6] R. D. Luce and A. D. Perry (1949). "A method of matrix analysis of group structure". Psychometrika 14 (1): 95–116.

[7] Freeman, L.C. (1977) "A set of measures of centrality based on betweenness,"Sociometry, 40:35-41.

[8] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) The PageRank Citation Ranking: Bringing Order to the Web

[9] American Society for Microbiology. "Genomics and social network analysis team up to solve disease outbreaks." ScienceDaily, 23 May 2011. Web. 24 Aug. 2011.

[10] Leskovec, J., Adamic, L. A., and Huberman, B. A. 2007. The dynamics of viral marketing. ACM Trans. Web, 1, 1, Article 5 (May 2007)

[11] Rogers, E. M. (1962). Diffusion of innovations. New York: Free Press

[12] Bass, Frank (1969). "A new product growth model for consumer durables". Management Science 15 (5): p215–227

[13] Krebs, Valdis.2001. "Mapping Networks of Terrorist Cells." 24(3):43-52.

[14] J. Travers and S. Milgram. An experimental study of the small world problem. Sociometry, 32(4):425–443, 1969.

[15] Jure Leskovec, Eric Horvitz, 2008, "Planetary-scale views on a large instant-messaging network" WWW '08: Proceeding of the 17th international conference on World Wide Web

[16] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," Phys. Rev. E, vol. 73, p. 016102, 2006.

[17] A. Clauset, C.R. Shalizi, and M.E.J. Newman, "Power-law distributions in empirical data" SIAM Review 51(4), 661-703 (2009)

[18] Ulrik Brandes, A Faster Algorithm for Betweenness Centrality, Journal of Mathematical Sociology 25(2):163-177, (2001).

[19] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000