Exploration into Socio-Economic Factors that can Potentially Affect Individual Internet Use Using Visual Analysis

CPSC 547: Project Proposal
November, 2017

Shirlett Hall, Peyvand Forouzandeh
smarionh@gmail.com, peyvand.forouzandeh@gmail.com

Contents

1. Introduction

Over the last 20 years, we have witnessed a far-reaching transformation of cities, due in part to technological revolution in terms of information and communication. The analysis of the information of the city must be placed within a global context since we cannot restrict ourselves to the physical boundaries of a society as it is organized and shaped around a system of networked communication technologies.[1]

The process of extensive demographic, social and economic concentration around cities and specifically Metropolitan areas are organized by series of hubs and are connected by telecommunications: the Internet and mobile communications. These regions are setting up around the world and create highly diverse regions in population density and activities.

2. Context of Use and Problem Description

The "Future Internet", offers the promise of efficiently supporting the needs of the information societies and digital economies for social and economic development. To achieve this objective, the Internet must be accessible, trusted and secure, as well as be able to robustly scale to meet the increasing reliance placed on it.

The Internet landscape includes users, who run applications over the Internet, Internet service providers, who sell public Internet access on a commercial basis, private network operators, policy-makers, intellectual property rights holders, and providers of content and services.

While efforts have been made previously to plot the topological structure in terms of the connections between Internet nodes–computer networks or Internet Service Providers, Internet maps (connectivity data) and visualizations to record the route of the information, most previous research and visualization efforts only considered the number of connections as an indicator of nodes.[3] While there are benefits in improving the efficiency of the Internet, improving the Internet in the future, both in the global south and global north, little attention was payed to the structure of the demand side and interlinks between socio-economic drivers of the users society around the world.

Governments and other policy makers are concerned with the gap in home Internet usage between people from metropolitan and rural areas. These research results highlight areas for further research and provide a basis for government agencies and industries to consider these associations in future policy formulation for regional development using ICT. These studies can lead to a profound understanding of future and current patterns in order to provide the best global ICT services as a tool for improving the quality of life. For instance, examining differences in home or individual Internet usage patterns between young and old, male and female, people in urban and rural areas, married and unmarried, well-educated and less educated, rich and poor, and employed and unemployed can uncover major differences in distribution levels.

Information visualizations targeting Internet penetration and geographic and spatial distributions of users as well as interactive maps of geospatial contents for various websites and

online networks has been of interest to data visualizers. However, we could not find many examples on the studies about the interlinks and connections between social and macroeconomic factors and the number of individuals using the internet in a global scale. This is the overarching goal of this project.

3. Personal Expertise

3.1. Shirlett Hall has a strong background in Financial Data Systems in both the Banking and Telecommunications industries. Currently, she is a Master's student in the Library and Information Science school with a specialization in Data Services. Specifically, she has been taking classes in Statistics, the use of R and Data Visualization. These classes and her experiences have provided her with many tools to 1. explore the datasets that include figures showing the use of the Internet and related socio-economic factors; and 2. to visually illustrate those figures. She still only has limited experience with both R and data visualization so the final project may only offer basic exploratory exercises. However, she will attempt to use as many resources as possible from the classes and from independent research to provide a good introduction to this topic.

3.2. Peyvand Forouzandeh has a strong background in design and is currently working as a PhD researcher with focus on urban technologies and smart cities. With statistics and social quantitative methods courses as well as her experience in information architecture and website design for local government decisionmakers and stakeholders (including urban planners, council members and general public) she has a profound understanding of the requirements and needs for visualization practices that best communicate with decision-makers. She wishes to use her skills and also use resources to find the best examples for visualizing the selected datasets and analyse the advantages and disadvantages of various visualization methods discussed in the class to be able to help with the provision of a set of recommendations for information visualization in this field.

4. Description of Domain, Task and Dataset

The domain, task, and dataset of this analysis project are described in the followings:

4.1. Domain

It has only been 25 years since Tim Berners-Lee made the 'World Wide Web' available to the public, but in that time, the Internet has already become an integral part of everyday life for most of the world's population. Additionally, almost two-thirds of the world's population now has a mobile phone.

Unlike most industries, which can be measured in terms of production or manufacturing, the Internet's economic contribution spans nearly every industry, and it is even creating entirely new industries driving innovation at a faster rate than we have ever seen in human history. The

lag between policy design and the innovation of the Internet and its businesses is understandable given the sheer speed with which the Internet has moved, but it is time to catch up.

Using one of the most recent datasets about Internet use in a global scale provided by the United Nations, we wish to present pictures of the locations of Internet users on national scales and find potential relations between Internet usage and country demographics including National Income per Capita, and access to electricity over an eight year span from 2008 to 2015.

By providing illustrations and graphs which enable us to compare socio-economic factors with Internet use, we aim to find answers to the following questions/issues:

1. Internet use and geographic boundaries
    - Are there specific relations or differences between the concentration of Internet use among geographic boundaries including multiple continents or regions?
2. Social factors and Internet use
    - Using visualizations that show the increase or decrease in Internet use over the terms of the 8 years, identify which social factors from the list selected show potential related effects on the growth and increased demand.
3. Future trends in socio-economic factors and Internet use
    - Can visualization help us understand future patterns and trends that connect socio-economic factors to the use of Internet in order to provide better services in countries with highest demand and need in the future years?
4. Global patterns: Global South and Global North
    - Is future growth in the total number of Internet users most likely to come from areas that are currently underrepresented?

We attempt to provide the basic information from the visualizations that we attempt to make in order to inform future research in specific areas related to our research topic.

4.2. Task

Due to the limited official data on Internet use, the rapid speed of the change in this area and time constraints, this analysis project can only answer parts of the questions posed in this proposal. Although this project is supposed to prove the benefit of information visualization by exploring more information to contribute to knowledge in this field, it is also to identify scenarios of using information visualization that improves the information extraction in the ICT domain for use in decision making process. The tasks of this analysis project are threefold:

- Selection of information visualization solutions

Analyze the characteristics of the datasets and analytic tasks related to domain questions

- Data analysis exploration

Basic analysis of datasets to find special patterns, relations and trends using Excel and R.

- Critique the current examples and suggest information visualization solutions in similar fields

Analyze pros and cons for each visualization method and approach with respect to analytic tasks, presentation and datasets. We will investigate into the different visualization techniques to be able to find best examples to present current state and future trends in our datasets.

- Conclusions

Based on the studies and analysis, with respect to the principles of information visualization learned in class, we will conduct suggestions and provide examples of visualization that we believe can best answer the questions and purposes of our study.

4.3. Dataset

The data was retrieved primarily from the United Nations Data Retrieval System. It is a central repository of data from other sources including the International Labor Organization (ILO), the World Health Organization (WHO) and the International Telecommunications Union (ITU). All four entities provide different components of the complete dataset used for this project. For example, two major attributes of our dataset is the number of Internet users per 100 inhabitants, and the number of mobile telephone subscribers per 100 inhabitants, on a country level. These attributes were found in separate tables within the dataset owned by ITU. The University of Groningen in the Netherlands also has an open dataset that provides a list of official languages for countries in the world.

The final dataset for this project is a compilation of thirteen attributes from five sources. With the exception of the language dataset, all individual datasets had two common fields. These are Country and Year. These two fields are used to merge all the individual datasets in R resulting in 1609 observations or items. The following Table 1 is a full description of the final dataset alongside a picture of the final dataset in Figure 1:

Table 1: Metadata Description of all Fields in the Factors Dataset

Field Name	Field Description	Data Type	Categorical, Quantitative, Ordinal	Number of Unique Levels/Range of Values
Country	Abbreviated and most recognizable name of each country. For example, the Democratic People's Republic of Korea is referred to as North Korea, and the Islamic Republic of Iran is referred to as Iran.	Character	Categorical	203
Year	Calendar Year that applies to the coverage period for an observation	Integer	Ordinal	2008 to 2015
Internet_Users_per_100	Number of Internet Users for every 100 inhabitants.	Double	Quantitative	0 to 100
Mobile_Subs_per_100	Number of Mobile Phone Subscribers for every 100 inhabitants.	Double	Quantitative	0 to 100
Tot_pop	Total Population	Double	Quantitative	-
Percent_rural	The percentage of all inhabitants that live in rural locations.	Double	Quantitative	0 to 1
Percent_urban	The percentage of all inhabitants that live in urban locations.	Double	Quantitative	0 to 1
GNI_per_cap	The Gross National Income per Capita	Double	Quantitative	-
Median_Life_Exp	The median life expectancy. This is an indication of the favorability of environmental	Double	Quantitative	-

	factors that persist in a population.			
Primary_Compl_Rate	The percentage of the population that complete a primary level of education.	Double	Quantitative	0 to 1
Per_Access_Electricity	The percentage of the population that have access to a source of electricity.	Double	Quantitative	0 to 1
Per_Adult_Unemployment	The percentage of adults 25 and over who are unemployed.	Double	Quantitative	0 to 1
Dom_Language	The first official language of the country.	Character	Categorical	86

Figure 1: Snapshot of the Factors Dataset

Country	Year	Internet_Users_per_100	Mobile_Subs_per_100	Tot_pop	Percent_rural	Percent_urban	GNI_per_cap	Median_Life_Exp	Primary_Compl_Rate	Per_Access_Electricity	Per_Adult_Unemployment	Dom_Language
Afghanistan	2008	1.840000	29.22037	23511400	77.32930	22.67070	370	58.22502	NA	42.40000	NA	Persian
Afghanistan	2009	3.550000	37.89494	23993500	77.04670	22.95330	470	58.60368	NA	47.88847	NA	Persian
Afghanistan	2010	4.000000	45.77817	24486600	76.76063	23.23937	520	58.97083	NA	42.70000	NA	Persian
Afghanistan	2011	5.000000	60.32632	24987700	76.47162	23.52838	570	59.32795	NA	61.51442	NA	Persian
Afghanistan	2012	5.454546	65.45219	25500100	76.17970	23.82030	720	59.67961	NA	69.10000	NA	Persian
Afghanistan	2013	5.900000	70.66136	26023100	75.88450	24.11550	730	60.02827	NA	75.15437	NA	Persian
Afghanistan	2014	7.000000	74.88284	26556754	75.58650	24.41350	670	60.37446	NA	89.50000	NA	Persian
Afghanistan	2015	8.260000	NA	NA	NA	NA	630	NA	NA	NA	NA	Persian

Based on a preliminary review of all the datasets, it is apparent that not all attributes are available for each country and/or year. This limits the findings that can be gained from visual analysis. We will either have to restrict the number of countries and/or years or attempt to complete the analysis with the acknowledgement that the conclusions are less than robust.

5. Proposed Infovis Solutions

5.1. Illustrations and hard drawn sketch

We will initiate our exploration with a choropleth map to illustrate the degree of Internet usage across the world. This can be done using the following libraries in R - rgeos, gpclib, maptools, and ggplot. Hue and depth of saturation can then be used to demonstrate the most recent state of penetration. These results will serve as the starting point for deeper exploration. We may either restrict the visual analysis to a specific world region or select the countries with the widest discrepancies in usage levels. Once this is completed, we will use the most relevant socio-economic factors, then select the best visual tools to identify the trends in the factors that

closely mirror the trends in Internet usage. The best tools should enable the user to perform exploratory data analysis in an efficient manner. The tools should also give the user the opportunity to filter or sort countries, and aggregate by factors such as income per capita. The user should also have a few options with the respect to the types of idioms. In addition to choropleth maps, we will use scatterplots and line graphs to support illustrations of trends. These options will also be dependent on the scenarios or the purpose of the exploration task.

5.2. Scenarios of use to address task

Scenario 1: If a user wanted to find out where and what the future trend is in Internet use for each country, they may be able to utilize a simple Exploratory Data Visualizer where clicking on the location of the country on a heatmap illustrating the Internet use, produces line graphs that compare socio-economic factors of that specific country with world average or regional rates. Below in Figure 2, we sketched an example of this scenario but the level of interactivity will depend on our skills in programming this map in R.

Figure 2: Sketch of the Choropleth Map

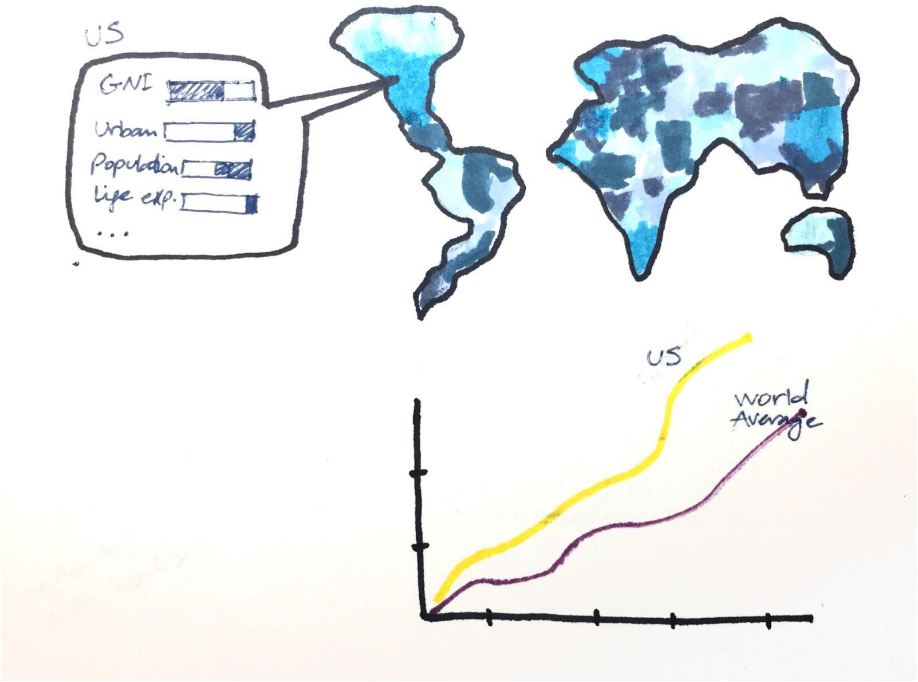
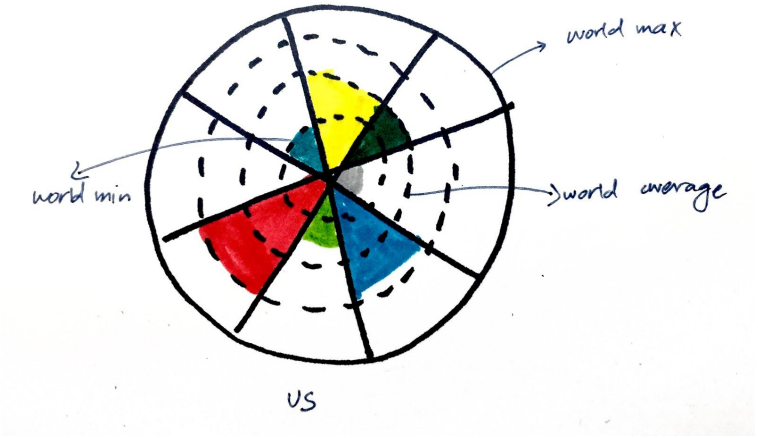


Figure 2 also shows an example of the accompanying illustration to show trends in Internet use for that country and world average rate in a trend graph.

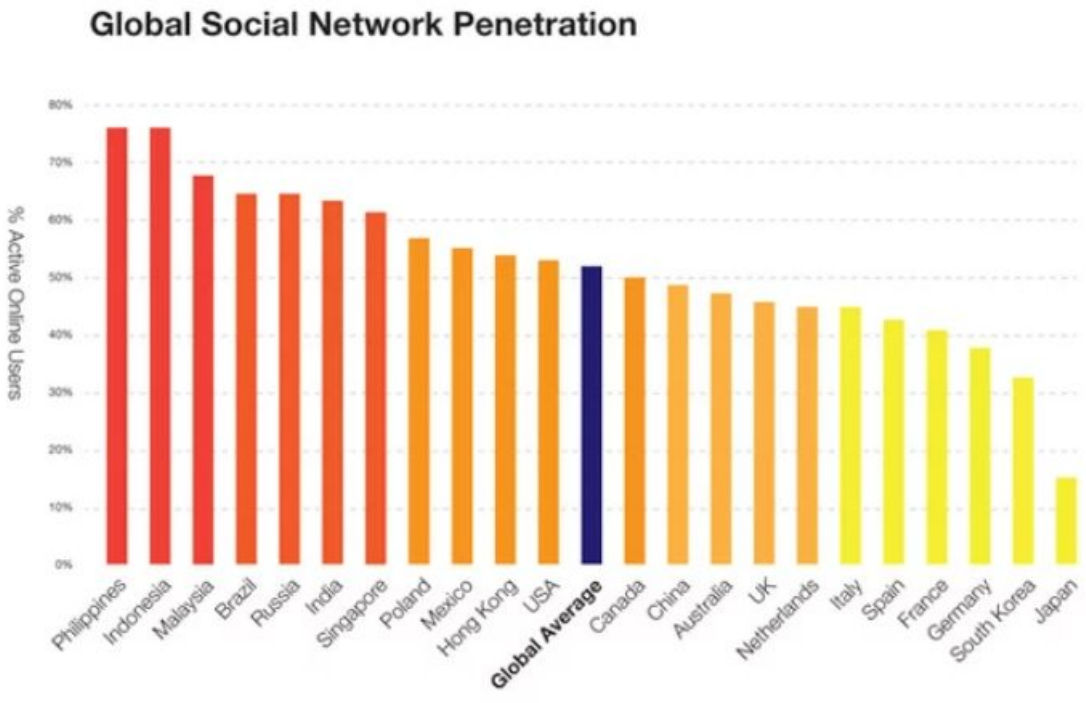
Scenario 2: Another scenario may involve displaying the countries which have the highest potential for investments in Internet in the future. In order to answer this question, we thought about two possible visualizations. One is a Polar area chart where we can see world maximum, world minimum and world average thresholds and multiple socio-economic factors affecting Internet use.

Figure 3: Sketch of a Polar Area Chart



The other possibility is a simple bar chart where we can see comparisons of Internet use among highly developing countries as shown in Figure 4.

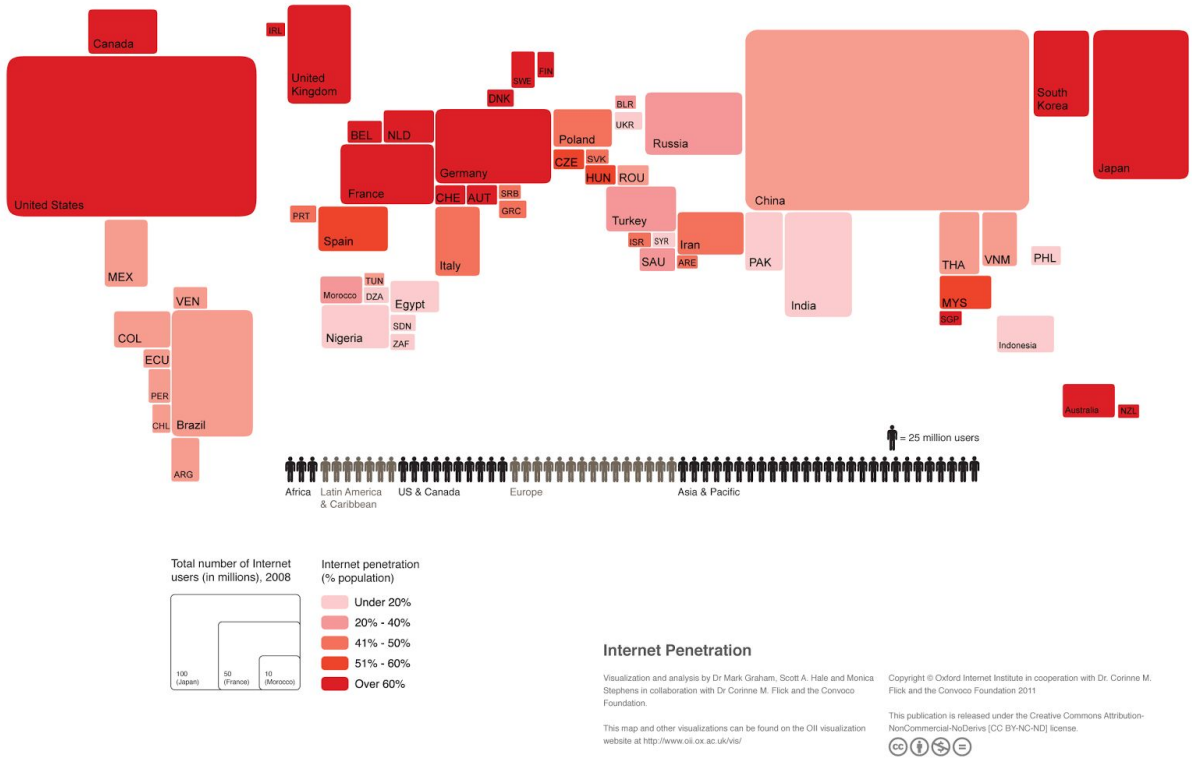
Figure 4: Bar Chart showing results of other Penetration Studies



Source: https://www.globalwebindex.net - https://www.buzzfeed.com/alexnaidus/cool-visualizations-that-show-how-we-use-the-internet?utm_term=.tsqomD0DW0#.vtjaXnVn0V

We could also find some other examples as in Figure 5, where the geographic locations are simplified by square shapes and the area of the square represents the the Internet penetration rate. We will explore the possibility of using R libraries to see whether this type of map can replace a heatmap and whether we have the capacity to develop a similar tool.

Figure 5: Diagram showing Internet Penetration Rates

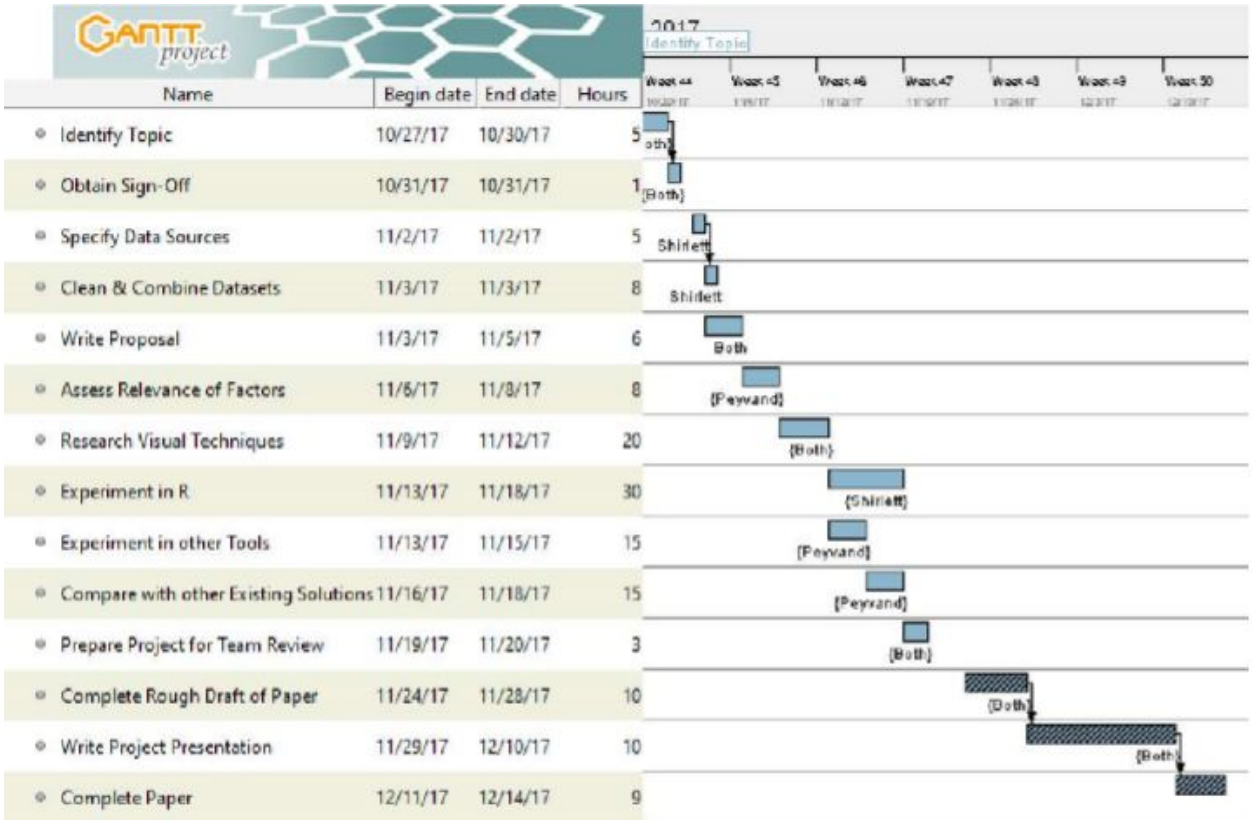


Source: http://geography.oii.ox.ac.uk/?page=internet-population-and-penetration-2008

6. Milestones

Figure 6 below shows the timeline of the milestones that need to be accomplished for completion of the project. It also displays the person assigned to tasks along with the expected number of hours associated with each task. It is estimated that the total number of hours that need to be allocated to the project is 140 hours spread evenly between the two group members.

Figure 6: Gantt Chart illustrating the Timeline for Project Completion



References

1. Castells, Manuel. “The Networked City.” The Social Fabric of the Networked City, First ed., EPFL Press, 2008, pp. V-xiii.
2. Pfieher, Geraldine et al. “The Social Fabric of the Networked City” First ed., EPFL Press, 2008, pp. V-xiii.
3. Graham-Row, Duncan. “Mapping the Internt” 19, June 2007. MIT Technology Review http://www.citationmachine.net/mla/cite-a-website
4. https://wearesocial.com/special-reports/digital-in-2017-global-overview
5. https://www.globalwebindex.net/
6. International Telecommunication Union "Individuals using the Internet 2005 to 2014", Key ICT indicators for developed and developing countries and the world (totals and percentage rates), International Telecommunication Union (ITU). Retrieved 25 May 2015.