Analyzing Mobile Games using a Social Network Analysis Approach

by

Nimat Onize Umar

Bachelor of Computer Science, University of Ilorin, Kwara, Nigeria, 2008

A Report Submitted in Partial Fulfillment

of the Requirements for the Degree of

Master of Computer Science

in the Graduate Academic Unit of Computer Science

Supervisor(s): Weichang Du, Ph.D., Computer Science

Donglei Du, Ph.D., Business Administration

Examining Board: Michael W. Fleming, Ph.D., Computer Science, Chair

Wei Song, Ph.D., Computer Science

This report is accepted by the

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

August, 2014

ABSTRACT

In recent times, analysis of user-generated data acquired from social media has proven to be beneficial in helping organizations make decisions about their businesses. This forms a basis for exploring other areas where social network analysis might be useful. In this report, I decided to look at the mobile games industry and see how accurate social network analysis can be, in making predictions of possible real world outcomes. Three approaches considered were how often a game is mentioned in social media (Frequency Count), sentiments attached to each game (Sentiment Analysis), and a game's position with other games in the network (Centrality Measures). Furthermore, using multiple linear regression analysis, five different predictive models were created by combining the three approaches. Finally, an evaluation of these approaches was done by performing correlation analysis between the rankings produced by each approach with the rankings in the Google Playstore. The best approach had a correlation coefficient of 0.58, which meant that the predictive ability of social network analysis for this industry is moderate.

DEDICATION

This report is dedicated to the Almighty God who through His infinite mercies gave me the grace to complete my master's program successfully.

Also to my parents, Mr. and Mrs. Umoru, for their continuous and total support throughout the period of my master's program.

ACKNOWLEDGEMENT

The successful completion of this report would not have been possible without the immense contributions of several people who deserve a mention as a mark of appreciation for all they did.

Firstly, I would like to express my profound gratitude to my supervisors, Dr. Weichang Du and Dr. Donglei Du for their direction and guidance while I was working on this report. I would also like to thank Dr. Michael W. Fleming and Dr. Wei Song for their input in helping improve this report.

My appreciation would not be complete without acknowledging all my colleagues and friends, whom I met here in Canada. Amongst many others, you are all appreciated; and I wish you all the best in your future endeavors.

Last but not least, I would like to thank my wonderful brothers, Mohammed B. Umoru and Abdulmuiz O. Umar for their words of encouragement throughout my master's degree program. May the binding love of the Almighty God continue to be with us (Amen).

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background

Social media is an online social platform where people interact and share content. There are several popular social media sites today, such as Twitter, Facebook, LinkedIn, YouTube and Blogger; they are fast becoming popular because many people spend a significant amount of time on the Internet and rely on its information for many aspects of their lives.

In recent times, analysis of user-generated data acquired from social media has proven to be beneficial in helping organizations make decisions about their businesses. This forms a basis for exploring other areas where social network analysis might be useful.

1.1.1 The Power of Social Networks

A network is basically entities (e.g. people, organizations, Internet, and financial entities) and their relationships [1]. Thus, we can say that there are several types of networks that are formed in our world, such as friendship networks, political networks, relationship networks, blog networks and food networks.

Networks are formed in the social media when people follow friends, add friends, comment on posts, reply to posts, retweet posts, etc. This data can be captured, analyzed,

and visualized to give more insight about the structure, size, and key positions of individuals in these networks [9].

There have been various breakthrough achievements so far, in the use of social network analysis techniques for viral marketing and making real-world outcome predictions. Areas that have seen such successes include box office predictions, election predictions, etc. Hence, organizations are beginning to see the importance of social media sites as an outlet to gather data for their business growth. One area that can fully take advantage of the free information available in social media is the mobile games industry.

1.1.2 The Mobile Games Industry

The mobile games industry is beginning to receive more attention as smartphones are becoming popular. It is one that relies heavily on its users, because the more users playing a game, the more friends of those users will likely be interested in downloading the game to play so that they can compete with each other or prove that they have a better score than their friends.

Over the last year, there has been a titanic annual growth of 130% in consumer spending on game apps, and consumer in-game apps spending has soared to more than double the size of spending on digital music [27].

Furthermore, according to gamesindustry.com, the mobile games market is expected to double in size by 2016 and reach US $23.9 billion [28]. Therefore, this area should not be overlooked in social network analysis, even though to date there has not been much focus on this industry.

1.2 Objective

The goal of this report was to investigate if an analysis of mobile game users' activities in social media could give a true reflection of what is happening in the real world (i.e. game's performance); if yes, the goal is to determine which approach can best predict a game's performance in the real world. This may then form a basis for future researchers to explore the social media as an outlet for making predictions of mobile games' performance in the real world. This investigation will determine whether social media is a valid source for such predictions.

This problem seemed ideal, because the mobile game industry is unique in the sense that regardless of how long a game application has been in the market, its success depends wholly on people (i.e. the number of users playing and downloading it), and how much attention it is receiving.

The focus was on free mobile games on the Android OS platform, as it is the platform in the mobile game enthusiasts' space that has one of the largest user-bases. Also, it was assumed that choosing this platform would allow a comparison of the results from my

experiments with actual results in the real world, since Google makes the daily rankings of all android free games available via the Google Playstore.

Although this may not cover the whole mobile game industry, the idea was that a study of this small area will allow more in-depth research and analysis which can be extended on other platforms for future research work.

1.3 Structure of the Report

This report comprises six chapters. Chapter 1 is the introductory chapter; it explains the motivation behind this report. Chapter 2 introduces the general concepts of graphs and networks, as well as giving deep insight into social network analysis. Chapter 3 focuses on the data that were used: the data source, how it was collected, and techniques used to prepare it for analysis. Chapter 4 describes how the mobile games network was formed and the different analyses performed on the data, such as frequency count, sentiment analysis, and centrality measures. Also, possible predictive models were defined by combining all three techniques using multiple linear regression analysis. Chapter 5 compares the results from the analyses described in Chapter 4 with the ranking in the Google Playstore using Correlation Analysis. Chapter 6 concludes the report and looks into areas that may be improved for future research.

Chapter 2. Literature Review

2.1 Graphs and Networks

Graph theory is the study of graphs, which are a mathematical representation of a network used to model pairwise relations between objects. This theory first started with Leonhard Euler in 1736, when he was asked to find a nice path across the seven Königsberg bridges. This problem required finding a walk through seven bridges in the city of Königsberg in Prussia that would cross each bridge once and only once. Since its initiation, graph theory has been applied to many fields of study such as chemical engineering, civil engineering, architecture, management and control, communication, operational research, combinatorial optimization, and computer science [14].

2.1.1 Fundamentals of Graph Theory

A graph is a pair $G = (V, E)$, which satisfies the expression $E \subseteq [V]^2$, where V is a set of vertices (or nodes) and E is a set of edges (or lines). It can be represented visually by drawing a circle for each vertex and joining two circles by a line if the corresponding vertices form an edge. A graph G with a set of vertices $V = \{v_1, \ldots, v_n\}$ is said to be a graph on V. The set of vertices of the graph is referred to as $V(G)$, and the set of edges as $E(G)$ [13]. Figure 2.1 shows a graph on $V\{1,\ldots,8\}$ with set of edges $E=\{\{1,2\}, \{2,6\}, \{2,8\}, \{3,5\}, \{3,8\}, \{4,7\}, \{4,8\}\}$.

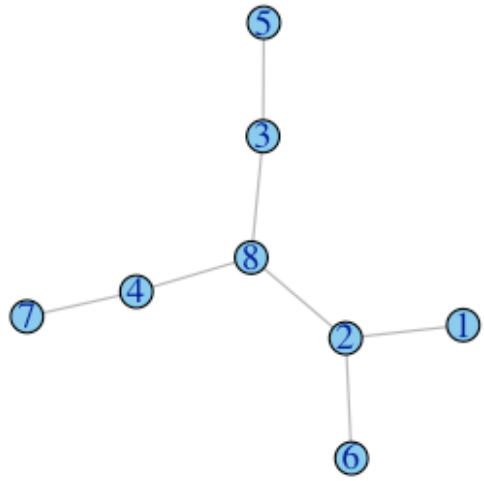


Figure 2.1 A Simple Graph

A graph may be simple, which means that it does not have any loops (vertex with connection to itself) or multiple edges, or a multigraph, if it has multiple edges. It may also be undirected, if there is no distinction between the two vertices associated with each edge, or directed, if its arc may be directed from one vertex to another. Lastly, it could be weighted if each edge of the graph has a number assigned to it, or unweighted, if there is no number assigned to each edge in the graph so that the weight on each edge is assumed to be the same [13][14][29].

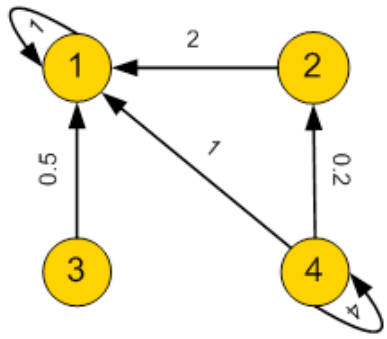


Figure 2.2 A Weighted Directed Graph [source: http://www.xatlantis.ch/]

In addition to representing graphs by listing the sets of connected vertices (Edge list), it may be represented using an adjacency matrix. This matrix shows which vertices (or nodes) of a graph are adjacent to which other vertices. The entry $A_{x,y}$ in the n-by-n matrix is equal to 1 if (x, y) is an edge, and it is equal to 0 if (x,y) is not an edge. This can be seen in the adjacency matrix and resulting graph in Figure 2.3:

$$
\begin{array}{c c}
 & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} &
\begin{pmatrix}
0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}
\end{array}
$$

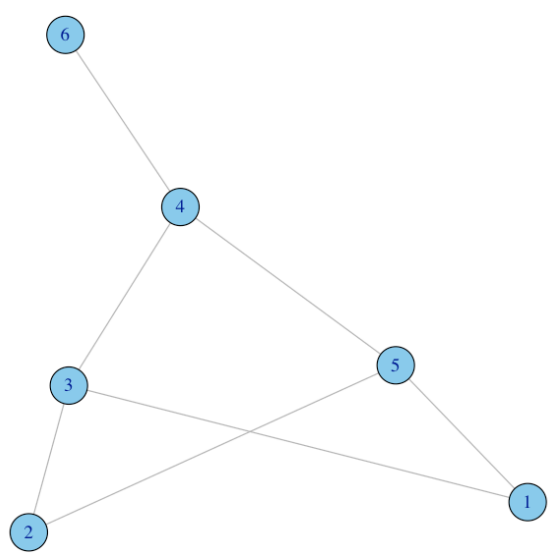


Figure 2.3 A Graphic Representation of the Adjacency Matrix

Below are some basic concepts and definitions in graph theory:

- Order: the order of a graph G is the number of vertices the graph has. It can be represented as $|V|$ [13].
- Adjacency: two vertices of a graph are said to be adjacent if these vertices are the end vertices of an edge [14].
- Incidence: an edge is said to be incident with a vertex if it is an end vertex of that edge. Two edges are said to be incident if they have a common end vertex [14].
- Walk: a walk w of G, is a finite sequence of vertices $w=\{v_0, e_1, v_1,...,e_k ,v_k\}$whose terms are alternately vertices v_i and edges e_i of G for $1 \leq i \leq k$, and v_{i-1} and v_i are the two ends of e_i [14].
- Trail: a trail t in G, is a walk in which no edge of G appears more than once [14].
- Path: a path p in G, is a trail in which no vertex appears more than once. The length of a path p_i denoted by $L(p_i)$ is determined by the number of its edges. p_i is called the shortest path between the two vertices v_0 and v_k, if for any other path p_j between these vertices $L(p_i) \leq L(p_j)$ [14].
- Distance: the distance between two vertices of a graph G is defined as the number of edges in the shortest path between the vertices [14].
- Connectedness: two vertices are connected in a graph G, if there exists a path between these vertices. A graph G is connected if all pairs of its vertices are connected [14].
- Cycles: a cycle is a path (v_0, e_1, v_1,...,e_k ,v_k) for which $v_0 = v_k$ and $k \geq 1$ (i.e. a cycle is a closed path) [14].

- Subgraph: a subgraph G_s of G is a graph for which $V(G_s) \subseteq V(G)$ and $E(G_s) \subseteq E(G)$, and each edge of G_s has the same ends as in G [13].
- Component: a component of a graph G is a maximal connected subgraph of G [13].
- Tree: a tree is a connected graph with no cycles [13].
- Forest: a forest is a graph where each connected component is a tree [13].

2.1.2 Network Properties

The structure of any network is dependent on its components (vertices and edges), such that if any of its components are altered, the overall behaviour of the structure will be changed. Similarly, if the location of its components changes, the properties of the structure will again be different. Therefore, in trying to discover unique properties of a network, the characteristics of its components would need to be looked at both individually and altogether [14]. Some standard concepts applied to define the properties of a network are the following.

Density: the density is the number of connections that are present in a network divided by the maximum number of possible connections that could exist in the network. The maximum possible number of connections in a network depends on the set of vertices V and on whether the network is undirected or directed. For an undirected network, the maximum possible number of connections it can have is $|V|(|V|-1)/2$; for a directed network it is $|V|(|V|-1)$. Therefore, the graph density for an undirected graph can be defined by the following equation:

$$density = \frac{2|E|}{|V|(|V|-1)}$$

where $|E|$ is the number of edges in the graph and $|V|$ is the number of vertices.

The density helps to define clusters. A cluster is a local region in a network with relatively high density and relatively few connections to other clusters [19]. The density of the undirected graph in Figure 2.3 is provided below:

$$density = \frac{2*6}{7(7-1)}$$

$$= 12/42$$

$$= 0.2857$$

Degree: the degree of a vertex is the number of edges connected to it, or the number of vertices adjacent to it. In social network analysis, the most popular vertices are those with the highest degree, signifying that they maintain the highest number of relationships. For a directed network, both the in-degree (i.e. the number of edges destined to vertex *v*) and the out-degree (i.e. the number of edges originated at vertex *v*) will be computed [16]. The degree for the undirected graph in Figure 2.3 is provided in Table 2.1:

Table 2.1 Degree for the Undirected Graph

Node	Degree
1	2
2	2
3	3
4	3
5	3
6	1

Weighted Degree: in weighted networks, the vertex with the highest degree is not necessarily the most popular vertex in the network because all its edges may be weak. On the other hand, a low degree vertex may be strongly attached to the network if all of its links are heavy. For this reason, the weighted degree of a vertex v is defined as the sum of the weights of all the edges connected to it [16]. The weighted degree for the weighted directed graph in Figure 2.2 is provided in Table 2.2:

Table 2.2 Weighted Degree for the Weighted Directed Graph

Node	Weighted In-Degree	Weighted Out-Degree
1	1+2+0.5+1 = 4.5	1
2	0.2	2
3	0	0.5
4	4	1+0.2+4 = 5.2

Degree Distribution: the degree distribution gives a frequency count of the occurrence of each degree (i.e. the probability of a vertex having a given degree). This reveals more information about the network than the degree of each vertex in the network because it provides essential information to understand the structure of a network [16]. The degree distribution for the undirected graph in Figure 2.3 is provided in Table 2.3:

Table 2.3 Degree Distribution for the Undirected Graph

Degree	Frequency
1	1/6
2	2/6
3	3/6

Clustering Coefficient: the clustering coefficient is the probability that any two neighbours of a vertex v (i.e. vertices directly connected to v) in a graph G are connected. Therefore, the clustering coefficient of the graph can be defined by the following equation:

$$clustering\ coefficient = \frac{2e_v}{k_v(k_v - 1)}$$

where k_v is the number of neighbours of v and e_v is the number of edges between them.

If the total number of neighbours of v is k_v, the maximum number of edges that can exist within that neighbourhood will therefore be, $k_v(k_v\text{-}1)/2$. Hence, the clustering coefficient

represents the fraction of the number of edges that are really in a neighbourhood. When the clustering coefficient is high for a vertex, that vertex is inciting its neighbours to interact with each other. Therefore, this measure can be considered as a measurement of the tendency of a given vertex to promote relationships among its neighbours [16]. The clustering coefficient for the undirected graph in Figure 2.3 is provided in Table 2.4:

Table 2.4 Clustering Coefficient for the Undirected Graph

Node	Clustering Coefficient
1	(2*0)/(2*(2-1)) = 0
2	(2*0)/(2*(2-1)) = 0
3	(2*0)/(3*(3-1)) = 0
4	(2*0)/(3*(3-1)) = 0
5	(2*0)/(3*(3-1)) = 0
6	NaN

2.2 Social Network Analysis

2.2.1 Overview of Social Network Analysis

Social network analysis is a technique that is only recently gaining popularity even though its application may be traced to as far back as the 1960s. A social network can be defined as a group of social actors that interrelate or exchange information with one another. The relationships under scrutiny could include friendship, common interest, kinship, dislike, influence, or, in the case of a scientific discipline, patterns of

communication or strength of association between members in a scientific community. A study of the interrelationships and communications between these actors provides interesting insights into how knowledge is spread throughout the community [15].

Mathematically, a social network is just a graph $G(V,E)$, where the vertices represent people, and an edge $(x, y) \in E$ denotes some type of social relationship between the people x and y [10]. In the social network analysis context, vertex will be replaced with node.

Social network analysis is the application of techniques, such as statistics, computer science, game theory, and/or graph theory in the study of relationships and ties between people, groups, organizations, and other entities. This area views social relationships in terms of network theory consisting of nodes and ties (also called edges, links, or connections) [17]. This analysis enables us to understand how networks are formed, the network structure, and how entities interact within the network. The purpose of social network analysis is to provide a way to quantitatively analyze relationships among these entities in a network. Furthermore, this analysis gives a much deeper level of understanding of the relationships, connections, and information flow within a network.

For instance, a smaller and strongly connected network will not be useful to its members. However, networks with lots of loose connections (weak ties) to individuals outside the main network will be useful. More open networks, with many weak ties and social connections, are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties. Thus, these open networks avoid the

same knowledge being shared continuously in the network [17]. Figure 2.4 is a simple social network showing friendships between a small group of friends:

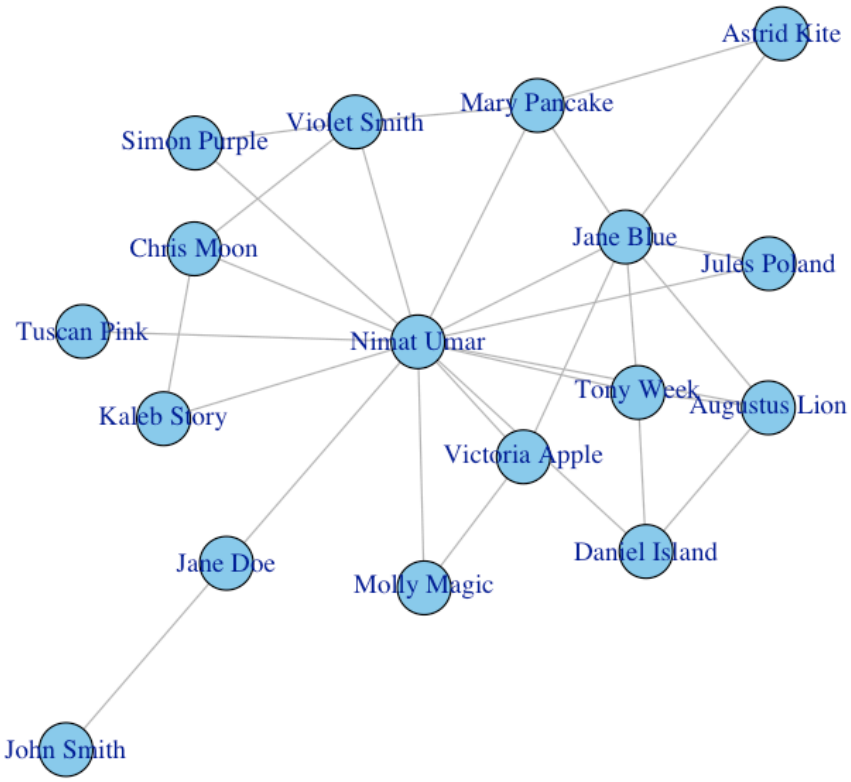


Figure 2.4 A Friendship Network

2.2.2 Why Social Network Analysis?

As discussed earlier, social network analysis techniques can be applied to study structures of any types of communications or relationships between actors in the network. Over the past few years, social network analysis has gained massive attention with its application in different academic fields such as computer science, economics, engineering, geography, information science, organizational studies, social psychology, and sociology.

Social network analysis is important because different types of networks are formed everywhere on a daily basis, and the amount of information that is constantly being shared and generated by actors in the network is increasing rapidly. Furthermore, these networks exhibit interesting phenomena such as the small world phenomenon or the six-degree of separation, the strength of weak ties, and the Giant component [2][10]. For example, the overall risk in the financial sector is an interesting area of study; social network analysis can help identify the roles that different financial entities or actors play in the financial network and may be extended to other entities.

Therefore, using social network analysis will provide an unambiguous, precise and explicit way to trace information flow and find influential actors in a network, which could help organizations in the following ways: gain feedback and insight on how to improve and advertise products better, tap into the wisdom of crowds to aid in making more informed decisions, draw up strategies to prepare for or avoid future problems, and find out how to strengthen or break connections or ties in the network [3]. Several applications of social network analysis that have seen such successes are:

- Combating terrorism: terrorist detection and terrorism prediction.
- Identifying the spread of diseases.
- Early detection of flu.
- Movie box office prediction.
- Stock market prediction.
- Sentiment analysis of Twitter.
- PageRank by Google.

2.2.3 Measures of Social Network Analysis

An important study in the area of social network analysis is determining influential actors in a social network. This can be determined by measuring the centrality of actors in a network relative to other actors in the network.

Centrality measure is a measure of influence, prestige, or relative importance of a node in a network. It answers the question, 'Who is the most important or central person in this network?' There are a vast number of different centrality measures that have been proposed over the years. The most common of them are Degree centrality, Closeness centrality, Betweenness centrality, Eigenvector centrality, and PageRank centrality [17].

These measures give us insight into the various roles that actors play in a network: who are the connectors, influencers, bridges, and isolates; where are the clusters and who is in them; and finally, who is in the core of the network? Hence, these measures are important since they reveal the importance and influence of all the nodes in the network [17].

2.2.3.1 Degree Centrality

In this measure, the node with the highest number of connections to other nodes is said to be the most central node in the network. For directed graphs, the node with the highest

outgoing connections is more central, while the node with the highest incoming connections is more prestigious [17].

Degree centrality of a node is different from its degree in the sense that it calculates the connections a node has with direct relation to the highest possible number of connections it can have in the network. Therefore, the degree centrality can be defined by the following equation:

$$degree\ centrality(v) = \frac{\deg ree(v)}{N-1}$$

where *degree*(*v)* is the degree of node *v* and *N* is the number of nodes in the network.

Nodes with high degree centrality are well positioned in the network because they have ties with many other nodes in the network and thus, can influence them. A central node occupies a strategic position in the network that serves as a source for larger volumes of information exchange and other resource transactions with other actors. The degree centrality for the friendship network in Figure 2.4 is provided in Table 2.5:

Table 2.5 Degree Centrality for the Friendship Network

Nimat Umar	Jane Blue	Violet Smith	Mary Pancake	Tony Week
0.8750	0.4375	0.2500	0.2500	0.2500
Augustus Lion	Daniel Island	Victoria Apple	Chris Moon	Jane Doe
0.2500	0.1875	0.1875	0.1875	0.1250
Jules Poland	Simon Purple	Kaleb Story	Molly Magic	Astrid Kite
0.1250	0.1250	0.1250	0.1250	0.1250
Tuscan Pink	John Smith			
0.0625	0.0625			

2.2.3.2 Betweenness Centrality

In this measure, the centrality is determined by the extent to which a node lies between other nodes in the network. This measure takes into account the connectivity of the node's neighbours, giving a higher value for nodes that act as a bridge along the shortest path between two other nodes. Therefore, the betweenness centrality can be defined by the following equation:

$$\textit{betweenness centrality}(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $s < t$, σ_{st} is the total number of the shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v [17].

Nodes with high betweenness centrality are able to control the flow of information or the exchange of resources in the network. Betweenness centrality is also an important measure to study the clusters in a network because it reflects the number of people to whom a person is connecting indirectly through their direct links [17]. The betweenness centrality for the friendship network in Figure 2.4 is provided in Table 2.6:

Table 2.6 Betweenness Centrality for the Friendship Network

Nimat Umar	Jane Doe	Jane Blue	Mary Pancake	Violet Smith
89.33	15.00	13.08	5.92	2.17
Victoria Apple	Tony Week	Augustus Lion	Chris Moon	Daniel Island
0.83	0.58	0.58	0.50	0.00
Tuscan Pink	Jules Poland	Simon Purple	Kaleb Story	Molly Magic
0.00	0.00	0.00	0.00	0.00
Astrid Kite	John Smith			
0.00	0.00			

2.2.3.3 Closeness Centrality

In this measure, the centrality is determined by how close a node is to all other nodes in the network. The closeness centrality for a node v is computed by simply finding the inverse of its farness [10][17]. The farness of a node v is the sum of its distances to all other nodes in the network. Therefore, the closeness centrality can be defined by the following equation:

$$closeness\ centrality(v) = \frac{1}{\sum_{i \neq v} d_{vi}}$$

where d_{vi} is the minimum distance from node v to node i (i.e., the sum of connections of all edges in the shortest path from v to i).

A node that is close to many others can quickly interact and communicate with them without going through many intermediaries. Therefore, nodes with high closeness centrality are able to spread information to other nodes faster in the network (i.e. the more central a node is, the lower its total distance to all other nodes.) [10][17]. The closeness centrality for the friendship network in Figure 2.4 is provided in Table 2.7:

Table 2.7 Closeness Centrality for the Friendship Network

Nimat Umar	Jane Blue	Violet Smith	Mary Pancake	Tony Week
0.0556	0.0385	0.0345	0.0345	0.0345
Augustus Lion	Victoria Apple	Jane Doe	Daniel Island	Chris Moon
0.0345	0.0333	0.0323	0.0323	0.0323
Jules Poland	Simon Purple	Kaleb Story	Molly Magic	Tuscan Pink
0.0323	0.0313	0.0313	0.0313	0.0303
Astrid Kite	John Smith			
0.0256	0.0217			

2.2.3.4 Eigenvector Centrality

In this measure, a node is said to be the most influential or central if it has the highest number of high quality connections. Scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. The eigenvector centrality algorithm follows this rule:

1. Start by assigning a centrality score of 1 to all nodes (v_i= 1 for all i in the network);
2. Recompute scores of each node as the weighted sum of centralities of all nodes in a node's neighbourhood;
3. Normalize v by dividing each value by the largest value;
4. Repeat steps 2 and 3 until the values of v stop changing.

Therefore, the eigenvector centrality can be defined by the following equation:

$$Ax = \lambda x$$

where A is the adjacency matrix of a graph G, λ is the largest non-negative eigenvalue of the adjacency matrix, and x is the corresponding eigenvector [2][21].

Nodes with high eigenvector are leaders of the network since they are connected to other nodes with mostly high ranks. An example of its application is Google's Page Rank

algorithm, which calculates websites' ranking based on links to them. The eigenvector centrality for the friendship network in Figure 2.4 is provided in Table 2.8:

Table 2.8 Eigenvector Centrality for the Friendship Network

Nimat Umar	Jane Blue	Augustus Lion	Tony Week	Mary Pancake
1.0000	0.6564	0.4977	0.4977	0.4442
Violet Smith	Daniel Island	Victoria Apple	Chris Moon	Jules Poland
0.3995	0.3903	0.3767	0.3244	0.3240
Simon Purple	Molly Magic	Kaleb Story	Astrid Kite	Jane Doe
0.2738	0.2692	0.2590	0.2153	0.2033
Tuscan Pink	John Smith			
0.1956	0.0397			

2.3 Sentiment Analysis

Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. A sentiment is a feeling of emotion or opinion that is expressed towards an object (i.e. an entity which could be a product, person, event, organization, or topic). Sentiment analysis involves using statistics, natural language processing, text mining, or machine learning methods to extract and identify the opinions, emotions, and sentiments expressed in texts. It is important to analyze a collection of texts rather than only one because one opinion only represents the subjective view of a single person, which is usually not significant [20][23].

The goal of sentiment analysis is to find out if the opinion or emotion expressed in the text is positive, negative, or neutral. This analysis is especially important when organizations are trying to get feedback from their customers about products or services they have delivered to them. Also, it may be applied when trying to find out the general opinions of people about an event that occurred, such as an election, a disaster, or a product that was released recently. In the past, organizations conducted surveys and opinion polls to get feedback about their products from customers. However, recently with the availability of user-generated content on the Internet, sentiment analysis is being explored for this purpose.

The first step in sentiment analysis is identifying the group of texts to be analyzed and its source. The source may be a product review blog, Twitter feed, or any other resources online. Next, the object on which the opinions were expressed will be determined. The texts may also be classified as subjective (i.e. expressing some personal feelings or beliefs) or objective (i.e. expressing factual information about an object). However, this is not necessary since some objective texts could also contain an opinion. Opinion sentences (i.e. a text that expresses a positive or negative opinion on an object) would then be extracted from the group of texts; these would mostly be subjective texts [20][22][23].

At this stage, the polarity (which indicates whether the opinion is positive, negative or neutral) of the opinion sentences will be calculated using a predetermined list of polarity

keywords. These keywords are adjectival or adverbial words and/or short phrases frequently found in texts that have been classified as positive or negative [20][22][23]. Finally, the results from the analysis will be summarized. Such a summary can also be visualized using a bar chart, a pie chart, or a word cloud. This way, a user or organization can easily see how existing users feel about a product.

Below is a fictitious example of a list of tweets made about an event:

1. This sucks. I can't stand this darkness #FrederictonPowerOutage
2. Had to throw out everything in my fridge. So angry right now!! #FrederictonPowerOutage
3. Does anyone know when power will be restored? #FrederictonPowerOutage
4. I think this is a good thing. For once we can stop relying on machines and spend more quality time with family #FrederictonPowerOutage
5. I've got power, send me a message if you need to charge your phones #FrederictonPowerOutage
6. Having dinner with friends tonight, this is the worst time for this to happen :(. #SadBirthday #FrederictonPowerOutage

In this example, the object on which the opinions were expressed was the Fredericton power outage. Tweets 1, 2, and 6 express negative emotions, while Tweet 4 expresses positive emotions. Therefore, the opinion sentences are Tweets 1, 2, 4, and 6. The result from the sentiment analysis of this event has been summarized visually in Figure 2.5. In the figure, the y-axis represents the strength of polarity and the x-axis represents the

positive, negative, and neutral emotions. Also, it can be observed that people had mostly negative emotions towards the power outage event.

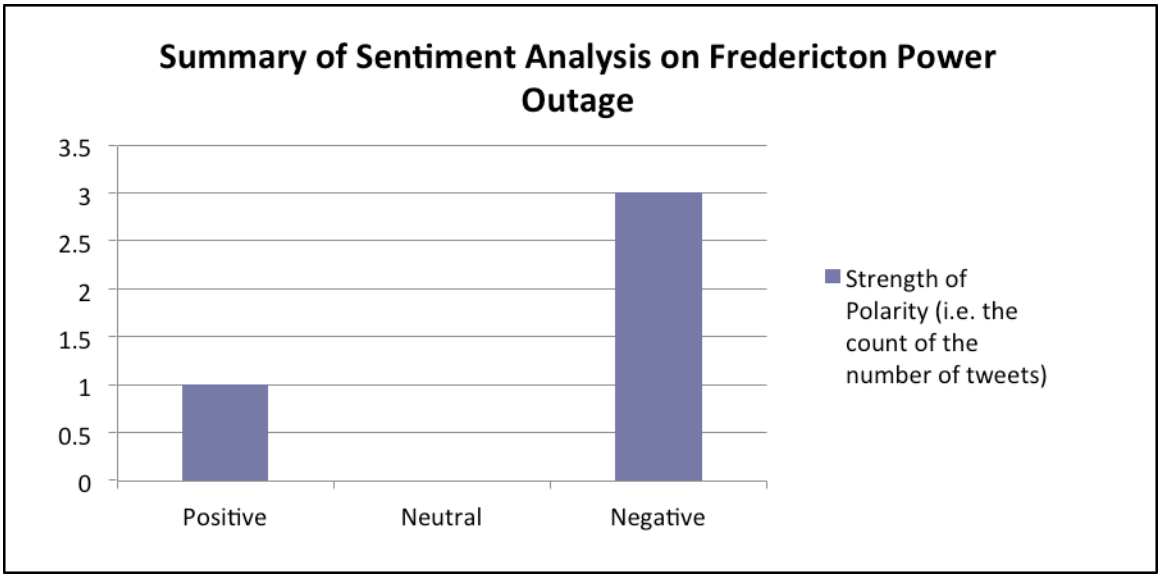


Figure 2.5 Bar Chart for Sentiment Analysis on Fredericton Power Outage

Chapter 3. Social Network Data

3.1 Identification of Data Source

The accuracy of results from any data analysis performed is largely dependent on the quality of the data used [12]. Therefore, it is important that a reliable data source is chosen for the area of interest. Social media data were ideal for this study because they are user generated; and since the outcome of a game's ranking in the market will be completely influenced by its users, the expectation was that these data would be close to accurate in capturing a real world outcome. Twitter was the best choice of all the social media sites currently available for the following reasons:

- Twitter is becoming one of the fastest growing social networking sites in the Internet because of its simplicity and the rapid flow and exchange of information within its space.

- It has been known to have a successful track record in making predictions, especially in the stock market and movie box office [18].

- It gives researchers access to public tweets through its Application Programming Interface (API). The search API allows researchers to retrieve past tweets by using criteria such as keywords and location, and the streaming API allows them to retrieve streams of new tweets using pre-set criteria as Twitter users initiate them [4].

- Networks are formed on Twitter. Take the mobile games industry for example; a network may be formed when a game is mentioned in a post, a post about a game is re-tweeted by a follower, or a friend replies to another friend about a game.

- People use Twitter as an avenue to talk about what games they are currently playing, the progress they are making in the games and what they think about the game. They usually do this with hashtags like #*NameOfGame* (e.g. #ClashOfClans for the game Clash Of Clans) included in their tweets. This makes it easier for other users that may not be among their followers to find these tweets through a hashtag search and possibly retweet them for friends in their own network, thus forming a network of tweets on the subject.

3.2 Data Capture

Data capture was performed using the streamR package on the R programming language (R-language), and Twitter streaming API. The initial goal was to choose the top 100 free games in the Google Playstore because there is really a limit to the number of games attracting attention in social media at any point in time; however, some mobile games were not chosen in order to ensure that ambiguous names were not used. For example, the game “Despicable Me” may also refer to a movie of the same name, and this would cause inaccuracy in the data collected. Eighty games out of the top 100 were thus selected for this report.

Next, data capture was initiated by modifying an R-language script originally written by Dr. Donglei Du and running it via the Twitter streaming API. The script contained keywords for the 80 games selected for this experiment, and it triggered a process that captured and saved real-time tweets containing a mention of one or more of the 80 games, for a period of 30 days. After capture was complete, 30 csv files containing a total of 1,330,635 tweets were produced.

3.3 Data Preparation

One of the most important steps in data analysis is the data preparation and cleaning stage. The data collected need to undergo some cleaning and manipulation before they will be able to produce useful information during analysis. This was done using the Python programming language, which is a powerful language for statistical programming.

The idea was to ensure that there were no inconsistencies in the data that were to be analyzed, check for errors in the data, correct errors detected, and transform the data into a form that would allow for easy analysis. Data preparation went through three different phases: Data Assessment, Data Reduction, and Data Transformation.

3.3.1 Data Assessment

In this stage, it was first important to understand the data being worked on by checking what items were available in each column, so I could know what useful information could be extracted from the data. It was noticed that the data contained extensive information, which could be used for different types of analysis. For example, information in the "country_code", "country", and "text" columns could give information about what games are popular in different countries or regions. The data were also checked for errors such as null values, outlier records, and invalid characters (e.g. *, ~, -, ", and _). Finally, errors identified were fixed. Some code from the Python script can be seen below:

```
#list all column headers in dataframe
sna_day6_v1.columns.values.tolist()

#display a summary of specified columns in dataframe
print sna_day6_v3["user_id_str"].describe()
print sna_day6_v3["in_reply_to_user_id_str"].describe()

#remove rows for user_id_str with empty values
sna_day6_v3=sna_day6_v2[pd.notnull(sna_day6_v2["user_id_str"])]

#define a function to cut of ".0" from a text
def cut_decimal(text):
    if ".0" in text:
        return text[:text.rfind(".0")]
    return text

#apply the function on all items of the stated column
sna_day6_v5. user_id_str = sna_day6_v5. user_id_str.apply(cut_decimal)
```

3.3.2 Data Reduction

This stage involved taking necessary actions to reduce the data, based on the assessment that was performed earlier, so that it could be easily manipulated. Since the focus of this report was on the performance of games in the real world, fields that were not relevant in determining the popularity and influence a game has in social media were removed. Finally, all outstanding errors that could not be fixed in the assessment stage were resolved. An example of one such error was ensuring that items in each column were converted into the correct data types that each column should hold. Some code from the Python script can be seen below:

```
#create a variable that holds the dataframe with only required columns
sna_day6_v2=sna_day6_v1[[“text”,“user_id_str”,“in_reply_to_user_id_str”
, “retweet_count”]]

#convert user_id_str column from float to string
sna_day6_v5. user_id_str = sna_day6_v5. user_id_str.astype(str)
```

3.3.3 Data Transformation

This stage involved converting the data into a format that could be read into R, from which a network could then be generated. Tweets in each row of the text field were initially reduced to only contain the keywords of each game. Furthermore, since users will mention games in different ways (e.g. Clumsy Bird, #ClumsyBird, clumsy bird, #clumsybird) in their tweets, these texts were unified to a single name format so that more than one node would not be created for the same game. Finally, a matrix was

created showing Twitter users' relationships with other users, and their affiliations with games in the network. Some code from the Python script can be seen below:

```
#create a mask for all keywords
clumsybird_mask  =  sna_day6_v1["text"].str.contains("clumsybird|clumsy
bird|\\#clumsybird|clumsbird|clumsy brd", case=False)

#search for keywords in text column of data frame and return the
keyword in a new column
sna_day6_v1.ix[clumsybird_mask,"clumsybird_header"] = "clumsybird"

#subset the dataframe into two rows with uniform columns
sna_day6_01 = sna_day6_v3[["user_id_str", "clumsybird_header"]]
sna_day6_01      =      sna_day6_01.rename(columns={"user_id_str":"A",
"clumsybird_header":"B"})

#create an edgelist from the matrix
sna_day6_app1 = sna_day6_usr.append(sna_day6_01, ignore_index = True)
```

Chapter 4: Mobile Games Analysis

4.1 Network Formation

4.1.1 Affiliation Networks

As discussed earlier, a network shows nodes and connections between those nodes, which represent relationships between them. However, things are a bit more complex in social networks because these nodes (representing people) may have connections with other types of nodes in the social network, called groups. Such representation is known as an affiliation network.

Affiliation networks belong to a class of graphs called bipartite graphs. A bipartite graph is a graph divided into two sets in such a way that every edge connects a node in one set to a node in the other set. In other words, there are no edges joining a pair of nodes that belong to the same set, and all edges go between the two sets. Bipartite graphs are very useful for representing data in which the items under study come in two categories, and an understanding of how the items in one category are associated with the items in the other is required [2].

An affiliation network is a bipartite graph that shows which individuals are affiliated with which groups or activities [2]. Affiliations in a social network play an important role in determining relationships between nodes that may not be obvious if affiliations were not considered, especially when we have more than one type of node in the network.

In the mobile games Twitter network, there are two types of nodes: Twitter users (game players) and mobile games. The relationship between two users will ideally be if there was some form of communication between them (i.e. a retweet, a favourite, or a reply). A game node on the other hand, will have a relationship with a user if the user mentions the game in his/her tweet signifying that the user is currently playing or has played the game.

The approach used was to first represent the twitter mobile game network in an affiliation matrix, with rows representing one type of node (users) and columns representing another type of node (mobile games). Subsequently, in order to accurately perform an analysis on the network, a single node relationship network needed to be generated from the dual node affiliation network, as in the sample affiliation matrix below:

$$\begin{array}{c} \\ 263637 \\ 728399 \\ 939400 \end{array} \begin{array}{c} \begin{array}{ccc} \text{clumsybird} & \text{candycrush} & \text{ironpants} \end{array} \\ \begin{pmatrix} 2 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

In the matrix above, we can see each Twitter user and their relationships with different mobile games. In this matrix, the focus was not only on the presence or absence of a relationship between the two nodes, but also on the strength of this relationship, with 0 signifying that there is no relationship between two nodes.

The affiliation matrix simply follows the rule that there are two types of nodes, one for game players and the other for mobile games. There will be a connection between a user node and a game node if the user mentions the game in his/her tweet. The strength of the connection will be determined by how many times the user mentioned the game in a day, thus forming a weighted affiliation network.

For example, the matrix point of "2" between “263637” and “clumsybird” means that a twitter user with id 263637 mentioned the game Clumsy Bird twice in a day. Thus, it can be concluded that the user has a stronger connection to Clumsy Bird than it has to Iron Pants since it has a matrix point of “1” with Iron Pants.

4.1.2 Dual Node Affiliation to Single Node Relationship

Ronald L. Breiger proposed an approach in his paper “The Duality of Persons and Groups” that could help in the analysis of the interpenetration of networks of persons and networks of groups that they belong to [11].

The general idea was that in a network containing both groups and individuals, the number of groups they belong to will define a tie between two individuals, and the ties between two groups will be the number of persons who belong to both groups, thus, forming two matrices: interpersonal ties matrix and intergroup ties matrix [11]. Therefore, the single-node game network will show that supposing a game user tweets

about game A and game B, his friend that is currently playing A would be motivated to try out game B too if he is currently enjoying game A.

Furthermore, if a game enthusiast performs a keyword search for a game, it will introduce him to other games that were contained in the same tweet. A typical example is the games Splashy Fish and Clumsy Bird; because of the similarities in the games, it was noticed that there was a trend with users who liked Clumsy Bird also liking Splashy Fish. This thus forms a basis for the relationship between the two games.

This concept also tries to prove that if a mobile game that is in the market is popular, there is a high chance that another game with common users will be popular as well. In this case, the conclusion is that we have formed a network where the relationship between two games, A and B, shows how likely a user that likes game A is to also like game B based on its number of common users. Below is a sample conversion of a dual node affiliation matrix to a single node relationship matrix:

$$\begin{array}{c} \\ 263637 \\ 728399 \\ 939400 \end{array} \begin{array}{c} \begin{array}{ccc} \text{clumsybird} & \text{candycrush} & \text{ironpants} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \end{array}$$

This will then be converted to:

$$\begin{array}{c} \\ \text{clumsybird} \\ \text{candycrush} \\ \text{ironpants} \end{array} \begin{array}{c} \begin{array}{ccc} \text{clumsybird} & \text{candycrush} & \text{ironpants} \end{array} \\ \begin{pmatrix} 2 & 2 & 0 \\ 2 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}$$

The diagonals show the frequency count for each game, and the numbers between two games are the number of twitter users that they have in common playing the game. A matrix point of 2 between Candy Crush and Clumsy Bird means that they both share two users.

4.1.3 The Mobile Games Twitter Network

The mobile games Twitter network had three components: a connected giant component and two other connected sub-components, all of which comprised a total of 80 nodes and 2886 edges. The giant component covered about two-thirds of the network, with 54 game nodes connected with each other; the other two sub-components had 24 and 2 game nodes connected with each other, respectively. From this observation, it could be assumed that since the games were connected with each other, information could flow freely within the network, and one game's position in the network could easily influence its ranking as well as other games' rankings in the real world. However, a game's position in the giant component will have no influence on another game's position in any of the other two sub-components, and vice versa. Therefore, information flow would only be

within a component, and there will be no flow of information from one component to another.

4.2 Frequency Count

The frequency count technique only focuses on a game's success based on how much it is being talked about in the social media, with the assumption that the fact that it is being talked about will make people interested in trying the game. A count of tweets containing keywords related to each game was performed and summarized using the Python programming language for all 80 games, and a ranking was assigned to each game based on its number of mentions. Therefore, the game with the highest number of tweets had the highest ranking, and the game with the lowest number of tweets had the lowest ranking. The top 10 mobile games' rankings derived from the application of the frequency count approach are provided in Table 4.1:

Table 4.1 Top 10 Mobile Games by Frequency Count

Rank	Game
1	csrracing
2	tetris
3	bingobash
4	ironpants
5	terraria
6	candycrush
7	angrybirdsstarwars
8	clashofclans
9	jetpackjoyride
10	Hayday

4.3 Sentiment Analysis

Sentiment analysis based on text classifiers was used to determine a user's positive or negative response to each game. The assumption was that a game with more positive tweets will encourage more people to download it and play it, and a game with more negative tweets will discourage people from downloading it. Therefore, we can quantify a game's performance based on its positive, negative, and neutral tweets.

The Sentiment analysis model used was based on an algorithm presented by Jeffrey Breen at the "Boston Predictive Analytics Meet Up" in 2011. Jeffrey Breen used this

algorithm to successfully determine the sentiment expressed in tweets about major U.S. airlines [32]. Thus, the same algorithm was applied on the Twitter data to determine the sentiment expressed in tweets about the different games, and classify each tweet into the positive, negative and neutral categories. The steps taken to achieve this were as follows:

- Extract text from the Twitter data and export to csv using the Python programming language
- Read csv files into R
- Load the sentiment word list on R[1]
- Remove special characters from the Twitter data
- Run Breen's algorithm on the Twitter data
- Group and summarize each game's sentiments into positive, negative, and neutral based on the scores that were assigned to each tweet using the Python programming language.

Next, in order to determine the ranking for each game, the following approaches were considered:

- Approach 1: positive tweets about a game will have twice as much effect in encouraging a user to download the game as neutral tweets. Also, negative tweets

[1] Over 6,000 words had been classified by Minqing Hu and Bing Liu into positive and negative sentiment words [26]. This classification used the Opinion Lexicon provided by Bing Liu, Minqing Hu and Junsheng Cheng [25].

about a game will have a great impact in discouraging a user from downloading the game. Therefore, the ranking for each game based on its sentiments can be determined by the following equation:

$$\frac{(2 * no.ofpositivetweets) \quad + \quad no.ofneutraltweets}{1 \quad + \quad no.ofnegativetweets}$$

- Approach 2: positive tweets about a game will have twice as much effect in encouraging a user to download the game as neutral tweets. However, negative tweets about a game will have no impact on a user's decision to download the game. Therefore, the ranking for each game based on its sentiments can be determined by the following equation:

$$(2*no.ofpositivetweets) \; + \; no.ofneutraltweets$$

- Approach 3: positive tweets about a game will have twice as much effect in encouraging a user to download the game as neutral tweets. Also, negative tweets about a game will have only a small impact in discouraging a user from downloading the game. Therefore, the ranking for each game based on its sentiments can be determined by the following equation:

$$((2*no.ofpositivetweets) \; + \; no.ofneutraltweets) \; - \; no.ofnegativetweets$$

- Approach 4: neutral and negative tweets about a game will have no impact on a user's decision to download the game. However, positive tweets about the game

will have an impact. Therefore, the ranking for each game based on its sentiments can be determined as follows:

$$no.ofpositivetweets$$

The top 10 mobile games' rankings derived from the application of the sentiment analysis approach are provided in Table 4.2:

Table 4.2 Top 10 Mobile Games by Sentiment Analysis

Rank	Approach 1	Approach 2	Approach 3	Approach 4
1	designthishome	csrracing	csrracing	csrracing
2	csrracing	tetris	tetris	bingobash
3	stickmandownhill	bingobash	bingobash	ironpants
4	doubledowncasino	ironpants	ironpants	tetris
5	clearvision3	terraria	terraria	designthishome
6	fruitninja	angrybirdsstarwars	angrybirdsstarwars	jetpackjoyride
7	jigty	candycrush	candycrush	clashofclans
8	ironpants	clashofclans	designthishome	angrybirdsstarwars
9	arcadeball	designthishome	hayday	terraria
10	doodlejump	jetpackjoyride	clashofclans	candycrush

4.4 Centrality Measures

The frequency count and sentiment analysis approaches were not sufficient for determining a game's performance because they both produced poor results after an

evaluation of both approaches was performed (see Section 5.1.1 and 5.1.2), so I decided to use the centrality measure approach to determine a game's ranking based on its position in the game's network that was constructed earlier:

- Weighted Degree: a game with the strongest and largest number of connections to other games will be easily discovered by other game enthusiasts that may not be aware of it; therefore, the most central node in this measure will have the highest ranking, and the least central will have the lowest ranking.

- Closeness: a game with the closest average distance (i.e. least number of intermediary nodes) to every other game in the network will have a greater chance of being noticed by users currently playing those games; therefore, the most central node in this measure will have the highest ranking, and the least central will have the lowest ranking.

- Betweenness: a game that lies between many other games in the network may play an important part in creating awareness for other games as well as for itself; therefore, the most central node in this measure will have the highest ranking, and the least central will have the lowest ranking.

- Eigenvector: a game that is connected to many popular games in the network will also eventually become popular; therefore, the most central node in this measure will have the highest ranking, and the least central will have the lowest ranking.

The top 10 mobile games' rankings derived from the application of the centrality measures approach are provided in Table 4.3:

Table 4.3 Top 10 Mobile Games by Centrality Measures

Rank	Weighted Degree	Betweenness	Closeness	Eigenvector
1	candycrush	candycrush	candycrush	candycrush
2	tetris	hayday	tetris	tetris
3	hayday	tetris	hayday	farmheroes
4	farmheroes	clashofclans	clashofclans	hayday
5	clashofclans	splashyfish	farmheroes	clashofclans
6	ironpants	ironpants	subwaysurfers	petrescuesaga
7	petrescuesaga	wordswithfriends	fruitninja	candyswipe
8	splashyfish	jetpackjoyride	templerun2	papapearsaga
9	papapearsaga	farmheroes	ironpants	fruitninja
10	fruitninja	fruitninja	dumbwaystodie	jellysplash

4.5 Regression Analysis

Regression analysis involves finding the relationship between two variables, one being independent and the other dependent. If the dependent variable increases as the independent variable increases, we know that there is a positive relationship between the variables. However, if the dependent variable decreases as the independent variable increases, we say that there is a negative relationship between the variables. Regression

analysis can also determine if there is no significant relationship between the independent and dependent variables [30][31].

There are different types of regression analysis that may be performed to find out the relationship between two variables. In this analysis, linear regression was used because the interest was in a linear relationship between the Google Playstore ranking and the result from the analyses.

In order to perform linear regression analysis, there must be some observed values for both variables. Then, a straight line that fits through all the points using the least square method will be determined. This line, which minimizes the errors when the estimated values are compared with the actual values, is called the line of best fit. The equation for the line of best fit is defined by the following:

$$Y = b_0 + b_1X$$

where Y = dependent variable, b_0 = y intercept, b_1 = slope of the line (negative if there is a negative relationship between the dependent and independent variables), and X = independent variable.

4.5.1 Multiple Linear Model

In order to combine more than one of the nine independent variables produced in the analyses performed, multiple linear models were derived using multiple linear regression [30]. The multiple linear model was defined by the following equation:

$$Y = b_0 + b_1X_1 + b_2X_2 \ldots b_9X_9$$

where Y = dependent variable, b_0 = y intercept, $b_{(1\ldots9)}$ = slope of the line on the nine independent variables, and $X_{(1\ldots9)}$ = independent variables.

Using the R programming language, five models were derived from applying multiple linear regression on the results from the analyses performed:

- Model 1: combining all approaches used in my analyses.

- Model 2: combining all centrality measures with the frequency count approach. Sentiment analysis was ignored because it produced the worst result overall after an evaluation was performed (see Section 5.1).

- Model 3: combining three centrality measures with the frequency count approach. Betweenness centrality was ignored because it produced the worst result out of all centrality measures after an evaluation was performed (see Section 5.1.3).

- Model 4: combining two centrality measures with the frequency count approach. Degree centrality was also ignored because it produced the second worst result out of all centrality measures after an evaluation was performed (see Section 5.1.3).

- Model 5: combining all centrality measures. This was done because the centrality measures approach produced the best results overall after an evaluation was performed (see Section 5.1).

The resulting equations of these models are provided in Table 4.4:

Table 4.4 Multiple Linear Regression Models for Mobile Games Prediction

Model No.	**Multiple Linear Model**	R^2	***p – value***
Model 1	PlaystoreRanking = 19.88648 + (-0.15731*Betweenness) + (0.65308*Closeness) + (0.10087*WeightedDegree) + (-0.21948*Eigenvector) + (0.59039*FrequencyCount) + (0.04678*Approach 1) + (0.04558*Approach 2) + (-0.19850*Approach 3) + (-0.41217*Approach 4)	0.3314	0.0005292
Model 2	PlaystoreRanking = 18.34515 + (-0.17284* Betweenness) + (0.64192*Closeness) + (0.18373*WeightedDegree) + (-0.25387*Eigenvector) + (0.08707*FrequencyCount)	0.2847	0.000123
Model 3	PlaystoreRanking = 20.29503 + (0.37531* Closeness) + (0.11634*WeightedDegree) + (-0.04070*Eigenvector) + (0.04073*FrequencyCount)	0.2698	8.432e-05
Model 4	PlaystoreRanking = 20.55874+ (0.38961* Closeness) + (0.04814*Eigenvector) + (0.05095*FrequencyCount)	0.2685	2.607e-05
Model 5	PlaystoreRanking = 20.0883 + (-0.1540*Betweenness) + (0.6687*Closeness) + (0.2089*WeightedDegree) + (-0.2726*Eigenvector)	0.2817	4.736e-05

The R^2 value (correlation of determination) in the table above gives the proportion of the variance of the dependent variable that is predictable from the independent variables, while the *p-value* is the probability of obtaining a result as close to the one that was actually observed, assuming that the null hypothesis is true [30][31][33].

The ranking of each game was determined by plugging in the values of the independent variables in the multiple linear model equations in Table 4.4. Thus, the top 10 mobile games' rankings derived from the application of regression analysis are provided in Table 4.5:

Table 4.5 Top 10 Mobile Games by Regression Models

Rank	Model 1	Model 2	Model 3	Model 4	Model 5
1	hillclimbracing	hillclimbracing	candycrush	candycrush	templerun2
2	subwaysurfers	tetris	tetris	tetris	Farmheroes
3	candycrush	bingobash	hayday	hayday	hillclimbracing
4	tetris	candycrush	clashofclans	clashofclans	subwaysurfers
5	templerun2	subwaysurfers	farmheroes	farmheroes	Tetris
6	hayday	hayday	ironpants	fruitninja	Candycrush
7	clashofclans	ironpants	subwaysurfers	subwaysurfers	Castleclash
8	farmheroes	clashofclans	fruitninja	ironpants	Clumsybird
9	ironpants	templerun2	templerun2	templerun2	Dragoncity
10	dumbwaystodie	dumbwaystodie	jetpackjoyride	jetpackjoyride	Bingobash

Chapter 5: Results and Evaluation

5.1 Correlation Analysis

Correlation analysis gives a statistical relationship between two variables or datasets. A correlation is said to be positive when one increases as the other one increases, and negative when one increases as the other one decreases. If they are on a line (so that one is a linear function of the other), the correlation coefficient is 1, and the value is 0 if unrelated [5].

Correlation analysis was performed on the results from the analyses performed in Chapter 4 in order to determine the technique that gives the closest linear relationship with the actual real world outcome. A scatterplot of Google Playstore rankings against rankings determined by the different approaches used was constructed (see Appendix III).

The Pearson's correlation method was the ideal correlation method for this experiment because it only looks for a linear relationship between two variables. Application of this correlation method showed how closely an increase in Google Playstore ranking corresponds with an increase in the result from the analyses. The strength of this correlation was determined by the value of the correlation coefficient (R).

Generally, all approaches had a low to moderate positive correlation with the Google Playstore ranking.

5.1.1 Frequency Count

The correlation between the frequency count approach and the Google Playstore rankings was low with a positive correlation coefficient of 0.3718. In addition to this, the breakdown of the number of games whose predicted rankings were within 0 to 10 spots of the Google Playstore rankings are provided in Table 5.1:

Table 5.1 Predicted Rankings Comparison: Frequency Count

Position in Rank	No. of Games
Exact Spot	1
Within 2 Spots	7
Within 4 Spots	13
Within 6 Spots	18
Within 8 Spots	21
Within 10 Spots	28

In order to determine if performing an analysis of mobile game users' activities in social media will not immediately predict the real world outcome, the rankings derived from applying the frequency count approach were compared with the Google Playstore rankings, one month and two months after completion of data capture. Thus, correlation analysis of the frequency count approach and the Google Playstore rankings one month and two months after data capture was again performed.

The result of this analysis became less accurate over time, with a correlation coefficient of 0.2963 for one month after data capture, and 0.2614 for two months after data capture.

5.1.2 Sentiment Analysis

The best result from performing a correlation analysis between the sentiment analysis approach and the Google Playstore rankings had a low positive correlation of 0.2500. This was even lower than the result derived from the frequency count approach. It was also noticed that there was a case of a low negative correlation. The results from performing a correlation analysis between the Google Playstore rankings and the sentiment analysis approach are provided in Table 5.2:

Table 5.2 Correlation Analysis: Google Playstore vs. Sentiment Analysis

Sentiment Analysis	Correlation Coefficient (After Capture)	Correlation Coefficient (1 Month After Capture)	Correlation Coefficient (2 Months After Capture)
Approach 1	-0.1143	-0.1711	-0.1539
Approach 2	0.2500	0.1684	0.1393
Approach 3	0.1984	0.0926	0.0687
Approach 4	0.1977	0.1522	0.1353

In addition to this, the breakdown of the number of games whose predicted rankings were within 0 to 10 spots of the Google Playstore rankings are provided in Table 5.3:

Table 5.3 Predicted Rankings Comparison: Sentiment Analysis

Position in Rank	No. of Games (Approach 1)	No. of Games (Approach 2)	No. of Games (Approach 3)	No. of Games (Approach 4)
Exact Spot	3	4	0	3
Within 2 Spots	5	8	6	10
Within 4 Spots	8	15	12	12
Within 6 Spots	9	20	14	15
Within 8 Spots	11	25	17	19
Within 10 Spots	13	28	21	23

From the results in Table 5.2, we can see that a game's sentiment does not have much influence on its performance in the real world. Upon investigation of why this might be, it was discovered that sometimes, negative sentiments expressed in tweets might not be about a dislike for a game but frustration over the difficulty level of the game; thus, this may not necessarily discourage other users from downloading the game but rather motivate them, since most mobile games enthusiasts love a good challenge. Furthermore, game enthusiasts use negative words as slang in tweets, which may not be negative in the context in which they were expressed. The game Flappy Bird is a typical example of how the sentiments expressed towards a game may not affect its performance. It had a lot of negative tweets from users about how frustratingly difficult it was to play; however, it performed very well in Google Playstore before it was taken down by its creator.

Recall that in the equations provided for Approaches 1 and 3, negative sentiments expressed in tweets about a game were assumed to have an impact in discouraging a user from downloading the game, with negative sentiments having a great impact in Approach 1 and only a small impact in Approach 3. This explains why Approach 1 had the lowest correlation coefficient of all Sentiment Analysis approaches. Furthermore, since Approach 4 only considered positive tweets, its low correlation coefficient compared to Approach 2 shows that neutral tweets also important in determining a game's performance.

Also, the accuracy of the result derived from performing a correlation analysis between the Google Playstore rankings and the sentiment analysis approach reduced over time, as in the frequency count approach.

5.1.3 Centrality Measures

Closeness centrality had the best predictive ability with a moderately positive correlation of 0.5169. The results from performing a correlation analysis between the Google Playstore rankings and the centrality measures approach are provided in Table 5.4:

Table 5.4 Correlation Analysis: Google Playstore vs. Centrality Measures

Centrality Measures	Correlation Coefficient (After Capture)	Correlation Coefficient (1 Month After Capture)	Correlation Coefficient (2 Months After Capture)
Betweenness	0.3344	0.1793	0.1419
Closeness	0.5169	0.4337	0.3784
W.Degree	0.4968	0.3989	0.3422
Eigenvector	0.5001	0.4738	0.4184

In addition to this, the breakdown of the number of games whose predicted rankings were within 0 to 10 spots of the Google Playstore rankings are provided in Table 5.5:

Table 5.5 Predicted Rankings Comparison: Centrality Measures

Position in Rank	No. of Games (Betweenness)	No. of Games (Closeness)	No. of Games (W.Degree)	No. of Games (Eigenvector)
Exact Spot	13	11	11	12
Within 2 Spots	15	24	16	18
Within 4 Spots	18	26	19	22
Within 6 Spots	20	27	23	25
Within 8 Spots	21	30	26	28
Within 10 Spots	23	33	34	31

From the results in Table 5.4, Betweenness centrality had the weakest predictive ability because a game with the highest Betweenness centrality only acts as a middleman for other games. Therefore, it may be promoting other games but not necessarily promoting itself. Closeness centrality had the best predictive ability because it will only take a short

time for mobile game enthusiasts to notice a game with the highest Closeness centrality since it is closest to more games in the network. Eigenvector centrality had the next highest correlation coefficient because unlike the Degree centrality that only considers the number of connections a game has; Eigenvector centrality also considers the quality of the connections. Therefore, a game that is connected to the highest number of games that are already famous in the industry will easily become famous too.

Lastly, just as it was with the frequency count and sentiment analysis approaches, the data in social media gave more accurate results in real time, but over time, they became less accurate.

5.1.4 Regression Models

The results from performing a correlation analysis between the Google Playstore rankings and the five regression models are provided in Table 5.6:

Table 5.6 Correlation Analysis: Google Playstore vs. Regression Models

Regression Models	Correlation Coefficient (After Capture)	Correlation Coefficient (1 Month After Capture)	Correlation Coefficient (2 Months After Capture)
Model 1	0.5756	0.5081	0.4491
Model 2	0.5336	0.4514	0.3953
Model 3	0.5194	0.4256	0.3696
Model 4	0.5181	0.4388	0.3836
Model 5	0.5307	0.4440	0.3873

In addition to this, the breakdown of the number of games whose predicted rankings were within 0 to 10 spots of the Google Playstore rankings are provided in Table 5.7:

Table 5.7 Predicted Rankings Comparison: Regression Models

Position in Rank	No. of Games (Model 1)	No. of Games (Model 2)	No. of Games (Model 3)	No. of Games (Model 4)	No. of Games (Model 5)
Exact Spot	2	2	3	2	2
Within 2 Spots	8	8	8	8	12
Within 4 Spots	15	17	16	14	16
Within 6 Spots	19	19	18	19	22
Within 8 Spots	24	23	23	22	27
Within 10 Spots	33	28	32	31	32

From the results in Table 5.6, it can be seen that combining more than one approach would help improve the predictive accuracy. Model 1 (combination of centrality measures, frequency count and sentiment analysis approaches) with a correlation coefficient of 0.5756, had the best predictive ability of all models. This seemed to be because the higher the number of variables used in the regression model, the higher the correlation coefficient. This can also be seen with Models 2 and 5, which had 0.5336 and 0.5307 respectively since all centrality measures were used. However, in Models 3 and 4, where some centrality measures were selected, the correlation coefficient reduced to 0.5194 and 0.5181 respectively.

5.2 Inference

From the analysis performed, several things can be inferred. For an industry like the mobile games industry, the predictive ability of social media is moderate. Upon investigation of this average performance by my analyses, it was discovered that two things may have caused these results:

- some mobile game enthusiasts may not yet be active Twitter users.
- new games introduced into the industry while data capture was in progress were not considered.

Of all three approaches used for my analysis, it was seen that the centrality measure approach gave the best result even though there was just a moderate positive correlation. This shows the importance of networks and relationships and how they influence a game's popularity over its frequency of mentions in tweets. Furthermore, it was determined that the sentiment analysis approach had little or no effect as a technique for mobile games prediction, because negative sentiments do not necessarily discourage users from playing or downloading a game; in fact, they sometimes may do the complete opposite. This can be seen from the very low correlation coefficient value for sentiment analysis techniques that included negative sentiments.

It was also observed that the accuracy of the rankings derived from the application of all techniques reduced over time. This means that, because of how volatile the industry is,

real time data would be ideal for predicting mobile games' rankings rather than historic data. Lastly, it was observed that using one method alone may not be sufficient, since a combination of all measures helped improve the result with the best correlation coefficient of 0.58.

Chapter 6. Conclusion and Future Work

6.1 Conclusion

In summary, first I looked at a game's popularity regardless of sentiments, which I called the frequency count approach. Next, I looked at the popularity based on how it may be affected by sentiments. Finally, I assessed a game's position in a network and how one game's position may affect another, by constructing a social network and computing the centrality measures.

The goal of this report was to determine the effectiveness of social media in predicting the rankings of mobile games. Although the best results had moderate predictive accuracies, this experiment revealed some interesting things about the industry. In the past, usage of social media for prediction focused a bit more on the tweet frequency. However, of all three approaches, centrality measures had the highest correlation coefficient values, even higher than the frequency count approach. This affirms the importance of using social network analysis to determine a node's position in a network and the relationship it has with other nodes in the network to predict real world outcomes.

Analysis of mobile games using social network analysis is an area that researchers should explore in the future. Although the best technique had a correlation coefficient of 0.58, it is hoped that this report will inspire an interest in this area since the expectation is that as more people (mobile games enthusiasts) start to embrace social media in the future, the result would become more accurate. Therefore, researchers should not be discouraged,

but see this as a place to start and an opportunity for future success, especially as reports show that this industry is expected to grow even more in the coming years [28].

On a final note, more awareness needs to be created about the strength of social media data, and how social network analysis can be explored to help take advantage of this strength. We are gradually moving towards an era where transactions in almost every aspect of our lives will be performed online, thus making a wealth of data available. After all, data is nothing without applying the right tools and techniques to transform it into useful information.

6.2 Future Work

This report only focused on the android platform with the expectation that an analysis of this small area may be extended to other platforms. Also, 80 of the top 100 games were isolated from other games in the industry such that new entrants into the market during data capture were not considered.

Therefore, other areas to look at for future research could be increasing the number of games used for analysis, extending the research to other platforms like iOS, Windows Phone and Blackberry.

One can also look at how specific users influence a game's performance by filtering out users with few followers; that is, assuming that a user with many followers will spread

the word about a game faster than a user with few followers. Finally, new games as they enter the industry should be considered while data capture is ongoing.

Bibliography

[1] Donglei Du, Bruce Spencer and Scott Buffett. Social Network Analysis Classes and Notes. *University of New Brunswick,* 2014.

[2] David Easley and Jon Kleinberg. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. *Cambridge University Press*, 2010.

[3] Jure Leskovec. Analytics and Predictive Models for Social Media: WWW 2011 Tutorial. *International World Wide Web Conference in Hyderabad, India*. Available at: http://snap.stanford.edu/proj/socmedia-www/ index.html#info, Tuesday, March 29, 2011.

[4] Lyric Doshia, Jonas Kraussabc, Stefan Nannabc and Peter Gloor. Predicting Movie Prices Through Dynamic Social Network Analysis. *Procedia - Social and Behavioral Sciences,* 2009.

[5] Michael A. Covington. Using R to Find Correlations. *Institute for Artificial Intelligence, The University of Georgia*. December 9, 2011.

[6] W. N. Venables, D. M. Smith and the R Core Team. An Introduction to R: A Programming Environment for Data Analysis and Graphics. *Version 3.0.2*. September 25, 2013.

[7] Kevin Sheppard. Introduction to Python for Econometrics, Statistics and Data Analysis. *University of Oxford*. February 24, 2014.

[8] Wes McKinney & PyData Development Team. pandas: powerful Python data analysis toolkit Release 0.13.1. February 03, 2014.

[9] How We Analyzed Twitter Social Media Networks with NodeXL. *Pew Research Centre, In association with the Social Media Research Foundation.* Available at:

http://www.pewinternet.org/files/2014/02/How-we-analyzed-Twitter-social-media-networks.pdf.

[10] David Liben-Nowell. An Algorithmic Approach to Social Networks. *Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology*, June 2005.

[11] Ronald L. Breiger. The Duality of Persons and Groups. *Social Forces, Vol. 53, No. 2, pp. 181-190,* December 1974.

[12] Oded Maimon and Lior Rokach. Data Mining and Knowledge Discovery Handbook. *2nd ed., DOI 10.1007/978-0-387-09823-4_2, New York: Springer Science+Business Media, LLC, pp. 19-31*, 2005.

[13] Reinhard Diestel. Graph Theory (Graduate Texts in Mathematics). *Springer-Verlag New York, vol.173.* 2000.

[14] Ali Kaveh. Structural Mechanics: Graph and Matrix Methods. *2nd ed. Taunton, Research Studies Press Ltd, New York etc., John Wiley & Sons Inc*. 1995.

[15] Diane M. Phillips and Jason Keith Phillips. A Social Network Analysis of Business Logistics and Transportation. *St Joseph's University, Philadelphia, Pennsylvania, USA and Ursinus College, Collegeville, Pennsylvania, USA*. May 1998.

[16] Luis López-Fernández, Gregorio Robles, Jesus M. Gonzalez-Barahona and Israel Herraiz. Applying Social Network Analysis Techniques to Community-Driven Libre Software Projects. *Int. J. of Information Technology and Web Engineering, 1(3), 27-48,* July-September 2006.

[17] Abhishek Awasthi. Clustering Algorithms for Anti-Money Laundering Using Graph Theory and Social Network Analysis. *University of Barcelona*, 2012.

[18] Sitaram Asur and Bernardo A. Huberman. Predicting the Future With Social Media. *Social Computing Lab, HP Labs, Palo Alto, California.*

[19] Bruce Hoppe and Claire Reinelt. Social network analysis and the evaluation of leadership networks. *The Leadership Quarterly 21 (2010) 600–619*.

[20] Bing Liu. Sentiment Analysis and Opinion Mining. *Morgan & Claypool Publishers*, May 2012.

[21] Cecilia Mascolo. Social and Technological Network Analysis. *University of Cambridge.* Cecilia Mascolo's Teaching. [online] Available at: http://www.cl.cam.ac.uk/~cm542/teaching/2011/stna2011.html [Accessed 18 Aug. 2014].

[22] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval. Vol. 2, No 1-2 (2008) 1–135.* 2008

[23] Bing Liu. Sentiment Analysis and Subjectivity. *Department of Computer Science, University of Illinois at Chicago.* 2010

[24] James Cook. Social Networks. *Sociology/Communications 375 at the University of Maine at Augusta,* 2014. COM/SOC 375: Social Networks at UMA, (2014). Syllabus. [online] Available at: http://www.umasocialmedia.com/socialnetworks/syllabus-spring-14/ [Accessed 18 Aug. 2014].

[25] Bing Liu, Minqing Hu and Junsheng Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of the 14th International World Wide Web conference (WWW-2005), Chiba, Japan,* May 10-14, 2005.

[26] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery*

and Data Mining (KDD-2004), Seattle, Washington, USA, pp.168-177, August 22-25, 2004.

[27] Tero Kuittinen. (2014). Mobile games have become ridiculously lucrative over the past year. [online] BGR. Available at: http://bgr.com/2014/02/19/mobile-games-spending-growth/ [Accessed 15 Aug. 2014], Feb 19, 2014.

[28] GamesIndustry.com. Mobile games market to double in size until 2016 and reach $23.9BN |Gamesindustry Blog. [online] Available at: http://www.gamesindustry.com/mobile-games-market-double-size-2016-reach-23-9bn/ [Accessed 15 Aug. 2014], 2013.

[29] Edward A. Bender and S. Gill Williamson. Lists, Decisions and Graphs Unit GT: Basic Concepts in Graph Theory. *University of California, San Diego,* 2010.

[30] Alan O. Sykes. An Introduction to Regression Analysis. *The Inaugural Coase Lecture, Vol. 20 of Working paper, Law School, University of Chicago, 1993.*

[31] Julian J. Faraway. Practical Regression and ANOVA using R. *University of Bath*, July 2002.

[32] Jeffrey Breen. slides from my R tutorial on Twitter text mining #rstats. [online] Available at: http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/ [Accessed 21 Aug. 2014], July 4, 2011.

[33] Roberts, D. (2014). Statistics 2 - Correlation Coefficient and Coefficient of Determination. [online] Mathbits.com. Available at: http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm [Accessed 26 Aug. 2014].

Appendices

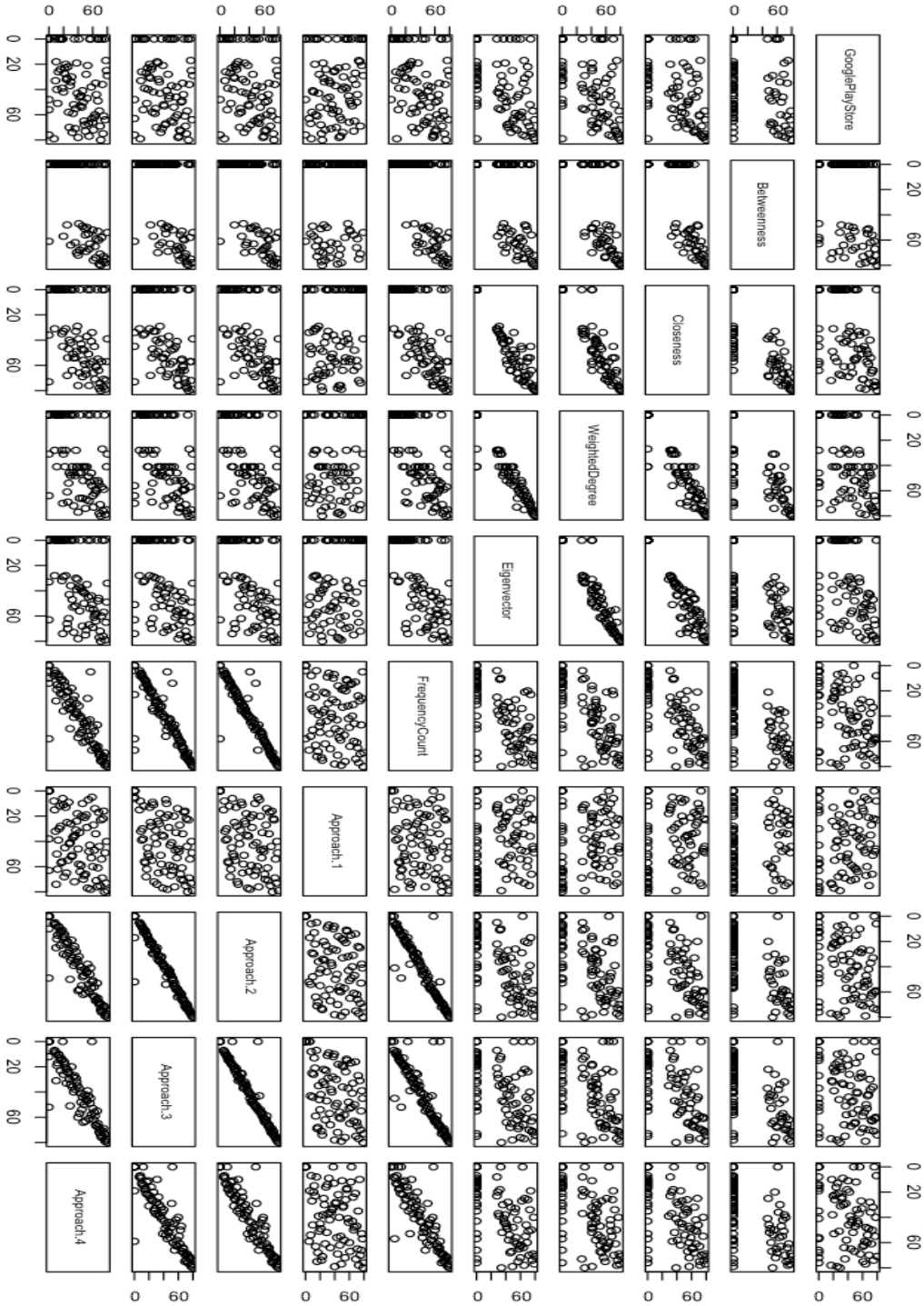
Appendix I: List of 80 Keywords Used for Twitter Data Capture

Game keywords	Game keywords	Game keywords
Candycrush	Zyngapoker	Ultimatepuzzle
Farmheroes	Angrybirdsstarwars	Stickmandownhill
Splashyfish	Wheresmywater	Splashythefish
Legocity	Bravefrontier	Theimpossibletest
Subwaysurfers	realracing3	Cookieclickers
templerun2	Ironpants	Doodlejump
Hillclimbracing	Icecreamjump	
Petrescuesaga	Csrracing	
Flowfree	Jellysplash	
Clashofclans	Bubblewitchsaga	
Fruitninja	Canyouescape	
Thesimpsonstappedout	Mahjongsolitaire	
Tetris	Cookieclickers	
Mytalkingtom	Angrybirdsfriends	
8ballpool	Empirefourkingdoms	
Castleclash	Wordswithfriends	
Dumbwaystodie	Bingobash	
Jetpackjoyride	Hungrysharkevolution	
4pics1word	Kidsdoodle	
100doors	Juicecubes	
Thesimsfreeplay	Highschoolstory	
Angrybirdsgo	Jigty	
plagueInc	Doubledowncasino	
Hayday	Petshopstory	
plantsvszombies2	Terraria	
Bejeweledblitz	Designthishome	
Clumsybird	Falloutbird	
Sonicdash	Dragoncity	
Jewelmania	clearvision3	
Papapearsaga	Blendoku	
Doodlejump	Arcadeball	
Bakerystory	Candyswipe	
Trafficracer	Slotmania	
Angrybirdsrio	Theimpossiblegame	
Glowhockey	Swipeoff	
deerhunter2014	Happypooflap	
Frozenfreefall	Tictactics	

Appendix II: Google Playstore Ranking at the End of Data Capture

Rank	Mobile Game	Rank	Mobile Game	Rank	Mobile Game
1	Candy Crush Saga	35	Clumsy Bird	69	Tank Riders2
2	Ultimate Puzzle	36	Traffic Racer	70	Design This Home
3	Farm Heroes Saga	37	Doodle Jump	71	High School Story
4	Subway Surfers	38	Zynga Poker	72	Blendoku
5	Temple Run 2	39	Angry Birds Rio	73	Stickman Downhill
6	Clash of Clans	40	Dragon City	74	The Impossible Test
7	Hill Climb Racing	41	100 Doors 2	75	Fallout Bird
8	Pet Rescue Saga	42	Glow Hockey	76	Tictactics
9	Splashy Fish™	43	Real Racing 3	77	Splashy The Fish
10	The Simpsons™: Tapped Out	44	Where's My Water? 2	78	Candy Swipe
11	Flow Free	45	Angry Birds Star Wars	79	Swipe Off
12	Fruit Ninja Free	46	Words With Friends Free	80	Terraria
13	Angry Birds Go!	47	Bakery Story: St Patrick's Day		
14	8 Ball Pool	48	Mahjong Solitaire		
15	Bejeweled Blitz	49	Kids Doodle - Color & Draw		
16	My Talking Tom	50	Juice Cubes		
17	TETRIS®	51	Hungry Shark Evolution		
18	Castle Clash	52	CSR Racing		
19	4 Pics 1 Word	53	Jelly Splash		
20	Jetpack Joyride	54	Ice Cream Jump		
21	Dumb Ways to Die	55	Empire: Four Kingdoms		
22	DEER HUNTER 2014	56	Bingo Bash		
23	Frozen Free Fall	57	Slotomania - FREE Slots		
24	Hay Day	58	Can You Escape		
25	Pet Shop Story	59	Bubble Witch Saga		
26	Plants vs. Zombies™ 2	60	DoubleDown Casino -		
27	Plague Inc.	61	Cookie Clickers™		
28	LEGO® City My City	62	Angry Birds Friends		
29	Papa Pear Saga	63	Jigty Jigsaw Puzzles		
30	Jewel Mania™	64	Iron Pants		
31	Sonic Dash	65	Clear Vision 3		
32	Brave Frontier	66	Arcade Ball		
33	Age of Warring Empire	67	Happy Poo Flap		
34	The Sims™ FreePlay	68	The Impossible Game		

Appendix III: Scatterplot of Google Playstore vs. Results from Analyses



Curriculum Vitae

Candidate's full name: Nimat Onize Umar

Universities attended: University of Ilorin, Kwara, Nigeria (2008, B.Sc)

University of New Brunswick, Fredericton (2014, MCS)