12-2013

Analysis of Social Networks in a Virtual World

Gregory Thomas Stafford
University of Arkansas, Fayetteville

Analysis of Social Networks in a Virtual World

Analysis of Social Networks in a Virtual World

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science

by

Gregory Thomas Stafford
University of Arkansas
Bachelor of Science in Computer Science, 2010

December 2013
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Dr. Susan Gauch
Thesis Director

Dr. John Gauch
Committee Member

Dr. Craig Thompson
Committee Member

ABSTRACT

As three-dimensional virtual environments become both more prevalent and more fragmented, studying how users are connected via their avatars and how they benefit from the virtual world community has become a significant area of research. An in-depth analysis of virtual world social networks is needed to evaluate how users interact in virtual worlds, to better understand the impact of avatar social networks on the virtual worlds, and to improve future online social networks.

Our current efforts are focused on building and exploring the social network aspects of virtual worlds. In this thesis, we build a social network of avatars based on their interaction in the Second Life virtual world and compare it to other social networking sites found on the web. Experimental results with data crawled from Second Life virtual worlds demonstrate that our approach was able to build a representative network of avatars in a virtual world from the sample data. The analysis comparison between virtual world social networks and others in the flat web allows us to gauge measures that better explore the relationship between locations linked by multiple users and their avatars. Using this comparison, we can also determine if techniques of personalized search and content recommendation are feasible for virtual world environments.

ACKNOWLEDGEMENTS

I would like to thank my family for their support and encouragement throughout my studies at the University of Arkansas as well as their understanding while I spent a majority of my time on this project.

Additionally I would like to thank Dr. Susan Gauch for the opportunity to work with her during this project and her guidance throughout my graduate education. I also thank my thesis committee, Dr. John Gauch and Dr. Craig Thompson, for reviewing my thesis.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Background

Virtual world environments allow users to navigate through and interact with online content in a 3D space using avatars as virtual representations. According to *Virtual Worlds Review*[1], virtual worlds such as Second Life, Open Simulator, and Active Worlds, are growing both in popularity and in number. In the last ten years, they have received significant attention from the public at large, from businesses and other organizations, and from scholars in disciplines as diverse as law, sociology, psychology, math, and, more recently, information systems [1].

Providing users with an easy way to locate online content quickly and efficiently has been a large area of interest. Several social networking web sites on ‘The Flat Web’ including Facebook, Flickr, LinkedIn, MySpace, Orkut and YouTube collect user information, user preferences, and online activities to better predict what web content will be of interest to users in the future.

Virtual world environments have the same need to quickly and easily locate content relevant to their interests, especially when content is presented with a full 3D environment that can be quite overwhelming to new users. Therefore, understanding how avatars behave when they connect to social networking virtual communities such as Second Life (see Figure 1) creates opportunities for better interface design, richer studies of social interactions, and improved design of content distribution systems.

[1] http://www.virtualworldsreview.com

In this work, we present an analysis of user workloads in both virtual worlds and online social networks. Comparing the social networks of virtual worlds to those found on the web not only shows whether or not current relevant content search is viable on the virtual world platform but may also elucidate how users behave differently in an immersive virtual environment compared to social networking sites. In addition, an analysis of the virtual world social network can give researchers insight to the behaviors of users when immersed in a virtual environment. Additionally, if a proper social network can be constructed, then techniques developed for the web that bring relevant content to users and generate user preference models may be implemented on this new platform. Bringing these technologies to virtual worlds will enhance this already expanding community and help new users find interesting content in an initially overwhelming environment.

Fig. 1 - Social interaction in Second Life virtual world

1.2 Motivation

Virtual environments allow users to navigate through online content in a three dimensional space. One of the most popular and widely known virtual worlds, Second Life, allows users to create avatars. An avatar, or a virtual representation of the user, is created to interact with and navigate the virtual world. Avatars may meet on any one of the numerous pieces of land and socialize, build, or conduct business. Many major companies actively use Second Life to host events or hold meetings. We will begin by a brief description of how Second Life allows for these interactions between their users.

1.2.1 Avatars

In order to be able to interact with Second Life and its users, a person may create a free account with Second Life on their website[1]. Each account is created with an avatar name and has one avatar associated with the account. Each user may create multiple accounts if they wish, enabling them to separate different personalities they wish to portray in the virtual world; one may be an outgoing socialite while another is a thriving storeowner. The appearance of an avatar can even be altered to whatever the user can design. An avatar can even appear as an object or animals if they wish.

The user logs into Second Life using a graphical viewer. Many exist but all operate similarly. The graphical viewer takes data sent from the Second Life servers and renders the in-game objects and avatars to be displayed to the user. The client facilitates the interactions between the user and the world allowing them to chat, design objects, write scripts to define the behaviors of objects, and more.

[1] http://www.secondlife.com

1.2.2 Groups

Once a user has created his avatar within Second Life, one natural next step is to find ways to locate other avatars to socialize with. There are many regions that are open to new avatars to help them get started and meet other people. Additionally user defined communities, or groups, help bring together avatars with similar interests. A group consists of two or more avatars and may have regions in which the group owns. The group may be public and allow members to join freely or private in which members may join by invitation only.

1.2.3 Link Definition

In social network analysis, it is important to understand how the entities of the network are related to one another. With social networking sites on the web it is easy to define these links from the explicit friendship requests between people. Depending on the implementation the links may be represented as directed links or undirected links. However, in Second Life, friendship link data between avatars is not available to be collected by our crawler. Therefore the links between nodes in the network must be defined in some other meaningful way.

Given the limited data we were able to collect using our crawler, we define our link edge between nodes by whether or not a pair of avatars subscribe to at least one group together. To generate a social network we were able to use the group membership list we created using a breadth first search on the publicly listed group subscriptions. The nodes of the networks are the avatars themselves while the undirected edge between nodes is defined as the number of groups that the avatars share in common. If the pair of avatars shares no groups between them, then no link exists, otherwise the weight of the edge is the number of groups in common. The more weight of an edge, the stronger the inferred relation between the pair of avatars. While we use no analysis requiring weights of an edge, it may be important for future research.

Other candidates for the link definition were picks in common and avatars seen in the same region by the crawler. Picks are Second Life's version of browser bookmarks that are user-defined locations of interest. The picks contain the addresses of these regions stored for quick access later. The crawler is able to acquire a full listing of an avatar's picks which have been used to generate hybrid preference models on a virtual world platform [2]. The data showing what avatars were located in the same area is limited given the behavior of the crawler. As there is a temporal and spatial aspect required for moving an automated avatar around a three-dimensional environment, the number of avatars seen in the same space is extremely sparse when compared to other link definitions. This limitation is not present with crawlers for the Web and provides a challenge should this type of relationship be desired.

1.2.4 Why Study Virtual World Social Networks?

New virtual worlds are appearing every day, while current implementations are growing rapidly. It is even said that an increasing portion of daily human interaction, economy, and culture will happen within these virtual environments [3]. As a transition to virtual worlds is seen, it becomes increasingly important to have efficient designs of how to present users with relevant information rapidly and effectively. Understanding the types of social networks that users of 3D virtual worlds create can help with this process.

With the additional requirement that two users must meet in the same spot at the same time within the virtual world, social interactions in 3D virtual worlds face unique constraints not seen on the Web. Social networking sites such as Facebook and YouTube allow users to post on one another's profiles or send messages to one another at any time. There is no requirement that both users be present at the computer for this fundamental interaction to take place. In Second Life, avatars may send messages to one another but it is not the main means of interaction.

Avatars are designed to been seen by one another and chat in front of one another, such as you would in most cases prefer speaking to a person directly rather than receiving mail from them. A crawler for social networking sites on the Web would be able to collect these basic interactions using HTML scraping techniques or an API if available, but, within the virtual world, collecting every avatar interaction with one another poses a problem. Therefore, whatever social network is constructed from the virtual world, it must be based on alternative link definitions until friendship or message data becomes available.

Analysis of the virtual world social network can give researchers insight to the behaviors of users when immersed in a virtual environment. Additionally, if a proper social network can be constructed, then techniques developed for the Web which bring relevant content to users and generate user preference models may be implemented on this new platform. Bringing these technologies to virtual worlds will enhance this already expanding community and help new users find interesting content in an initially overwhelming environment.

Studying the social network of virtual environments also has far reaching impact. Comparing the social networks of virtual worlds to those found on the Web not only shows whether or not current relevant content search is viable on the virtual world platform but may also elucidate how users behave differently in an immersive virtual environment compared to social networking sites. Sociology researchers may find interest in the social network analysis as a whole. Clustering of users and their changes over time could reveal information regarding if users actually become fully involved in the virtual world or tend to keep to themselves.

To enhance the attractiveness of virtual worlds like Second Life, we attempt to solve the fundamental information retrieval problem of providing users with useful and interesting content. With the superfluence of data and content in this virtual world we investigate if such user content recommendation systems are viable on this platform by analyzing the social networks generated from data crawled. Using these virtual world social networks we compare with social networks found on sites like Facebook and LiveJournal to see if the structure and properties are similar. As social networks are the foundation for many of these recommendation systems, we can compare these social networks with networks found from popular social sites to can gain insight as to whether or not similar systems can be implemented successfully on virtual world platforms.

1.3 Organization of this Thesis

In Chapter 2, we present a summary of related work on social networking in virtual worlds. Chapter 3 introduces how we have built a social network of avatars and present measures to compare the social network of virtual worlds to those on the current flat web. In Chapter 4, we report our experimental results and evaluation of this comparison. Finally, in Chapter 5, we present conclusions and discuss our ongoing and future work in this area.

2. RELATED WORK

This chapter surveys related research on virtual worlds, social networks, and social network analysis.

2.1 Virtual Worlds

Because virtual worlds are a relatively new and evolving platform for online interaction, there has only been a limited amount of research done to evaluate the effect of these worlds on day-to-day human life. We discuss this research and also summarize work on the behavioral impact of virtual worlds, current architectures, and virtual world crawling techniques.

2.1.1 Impact of Virtual Worlds

To enrich the scientific community's knowledge regarding virtual worlds, Bainbridge has taken a close look into the impact of two specific and extremely popular virtual worlds, World of Warcraft and Second Life [3]. Each virtual world has its own attributes that may be explored for research on virtual worlds. In addition, both can contribute to significant studies in other scientific fields. For example, both games exhibit in-world economies in which the prices of items or services in the world fluctuate based on supply or demand. Second Life's economy may even be transferred into real world money depending on specific exchange rates. Additionally, the demographics of these worlds may be brought to light; the majority of users in these virtual worlds were mostly male but were across many various demographics and occupations.

Bainbridge explains that the significance of virtual worlds to the scientific community is a direct result of the sheer amount of observable data inherent in these virtual worlds [3]. Second Life has provided in game tools for users, who may pay a small fee, to set up experimental labs where other real users can come by and take part in surveys, quickly increasing the sample group. Epidemics have been reported to spread across World of Warcraft due to the fact that users are frequently in proximity of one another. In team events, such as World of Warcraft player-versus-player combat, data regarding player collaboration to accomplish a goal is available and may be studied. Each of these has an obvious amount of data that may be collected and analyzed by any willing researcher.

The research opportunities available cover many disciplines. Social scientists will find interest in the behavioral aspects of the virtual world: transferal of behavioral norms from real world to virtual world, cultural boundaries, cooperation among users and even virtual world addiction. Economists can easily collect mass information regarding World of Warcraft's demographics and item pricings using their accessible add-on system. Information and computer sciences will find a reason to push current technology to its limits to provide a world that is as seamless as possible, ways to improve artificial intelligence to provide a more engaging world for the users, and as these worlds develop virtual reality systems have a potential to improve in similar ways.

In response to the variable research potential of these virtual worlds, and maybe many others, it is anticipated that virtual worlds will continue to grow as well as scientific interest. Reports on how and why users personalize avatars and the resulting social interactions have yet to be explored. Human response to ever-improving artificial intelligence and the cognitive process that shape the users behaviors may also be extracted from these worlds [3]. It is apparent that the research potential will grow as virtual worlds increase in complexity and popularity.

Parallel to social and economic research opportunities, Collins elaborates on the potential impact that virtual worlds may have on business and education [4]. He discusses the fact that virtual reality has played a large role in entertainment media in the past and with the development of virtual worlds comes the increasing potential for virtual reality to become part of daily life. While virtual worlds bring the technology to make virtual reality an actuality their impact on the education and business is becoming more apparent. The increased connectivity between distanced parties and the change in ways a person can interact with information are having a dramatic effect on everyday interaction.

Opportunities for businesses to expand their empire across multiple platforms have been taken advantage of by several companies. IBM and Intel both have reserved virtual spaces on the Second Life servers to exploit easy to use interaction channels to bring distant offices closer. The author states that these companies are simply preparing for the future and anticipating the evolving technology and adjusting the company's exchanges to various platforms. These large companies investing into virtual worlds makes it apparent that there are certainly advantages of using virtual worlds, and the article articulates the benefits of hosting a virtual campus in these virtual worlds to provide a platform for students, faculty, and staff interaction.

Unfortunately, there are a few down sides to this evolving technology. Travel between various virtual worlds is not possible for everyday users. Typically, virtual worlds do not provide mechanisms for an avatar to transfer its properties between different companies' servers; no standardized methods exist to facilitate such a transfer of data. Additionally, some virtual worlds with dynamic content require expensive top-of-the-line equipment to operate the virtual world client in a way that is acceptable by users. Despite these issues and with the advancement of technology, the article supposes that the advantages of having a virtual world presence will be hard to overlook in the upcoming years.

2.1.2. Behavioral Effects

A user's psychological patterns are a topic of interest to this research as this research seeks to find behavioral patterns of the entities involved in the virtual world and the users behind them. A question regarding the difference between a user's online and offline persona has be asked by many researchers and Ducheneaut et. al. attempt to tackle this question using virtual worlds and the avatars that the users control [5]. An avatar is a digital body that the user may customize within the constraints of the application; a user may wish to create any identity with this customization. In their study, approximately two hundred subscribers of various virtual worlds are asked to outline their avatar appearance preferences and some questions regarding demographic information.

The research team finds a range of interesting results from their experiment. Overall, more users were male than female and the mean age for each game varied greatly. Furthermore, hairstyle and color were considered the most important avatar appearance attributes. While varying demographics showed different approaches to avatar creation, users almost always modified their appearance and in most cases the changes were deemed favorable according to Western cultural norms, suggesting that many virtual world users seek some sort of perfection in their appearance. More interestingly, while physical changes were nearly ubiquitous, psychological differences were extremely minimal if existing at all. These findings provide an important insight to the mind of the user that should be considered when analyzing results of comparable research.

Doodson performs a similar survey technique in the Second Life virtual world [6]. The survey consisted of a Five Factor Model set of questions each answerable by a one-to-seven scale of agreeableness. The survey was performed with a sample group of over one hundred participants. Contrary to the previous research, it was found that although offline personalities were a good predictor of virtual world personalities, the two are not similar. The approaches to the experiment were similar in both studies but perhaps the relatively small sample in both cases allowed for such a fluctuation in personality expectations of offline and online users. Whatever the case, it is important to realize that behavioral patterns and personalities in and out of a virtual world may differ and relating findings within a virtual world to the real world may have its own complications.

2.1.3 Architectures

In the examination of virtual world architectures, Thompson discusses how virtual worlds have grown to what they are today and what we may expect to see in the future in terms of standards, interoperability, and growth [7]. Virtual worlds immerse the user into a 3D environment and use it as a means to interact with others or explore what the world has to provide. The most popular virtual world, Second Life, provides this service through a graphical client that pulls data from servers containing the rich content contained within this world. This type of architecture differentiates from gaming platforms as it is impractical for providing fast paced content, but it is suitable for social and exploratory purposes. Thompson also discusses how virtual worlds converge at a higher level in their architectures and how this may be exploited to provide end users with a seamless transition between worlds. A standard may be able to be constructed to integrate these worlds and allow users to switch worlds as easily as it is to browse multiple web pages. This standard would then support the growth and popularity of virtual worlds providing users with more dynamic and 3D web content.

Benford et al. note that supporting virtual worlds that are densely populated with rich content and interaction is a very ambitious goal. One of the main problems is the scalability of such a platform [8]. With an almost guaranteed high volume of traffic, a virtual world's success will be greatly impacted by network bottlenecks and end-user hardware limitation. Many existing large-scale virtual worlds attempt to counter this problem by dividing the world into regions or parcels. These individual environments mimic the world as a whole but allow the server load to be distributed across machines. Among the architectures discussed is the classic client/server architecture with one common server. One advantage of this architecture is that the developers may tailor the environment to the average client’s hardware for a network/hardware capacity based tradeoff. Additionally there are two types of peer-to-peer architectures, unicast and multicast. In unicast architectures clients send information directly to other clients, bypassing servers to avoid high server loads; the drawback of this system implementation is an intensive demand on bandwidth. Multicast architectures broadcast information to many other clients while avoiding the load on the network by using bandwidth-efficient protocols such as IP multicast.

Kazman has observed the current limitations of virtual world architectures and has proposed the WAVES architecture [9]. In virtual environments where there are many actors, such as flight simulators, point-to-point, broadcast, and hierarchal architectures are limited by communication bandwidth and bottlenecking between processes. The WAVES architecture proposes a network of message managers which individually govern multiple hosts, who are in turn responsible for input and output to the environment. To avoid the overhead found in classic programming practices such as inheritance and polymorphism, an objoid, or encapsulation of state, is used to notify the entities of the network of change in predicted behavior. Although the success of this system is based on the predictability of the environment, the notion of the objoid allows for a more lightweight system of communication between several entities in the virtual world. While the popularity of this architecture may not be widespread, it is important to note that virtual environments are not limited to current architectures and may evolve in their continued use and growth.

2.1.4 Crawling

Varbello et al. produced a paper closely related to this research that elucidated the approach the authors took on ‘spidering’ Second Life [10]. For the crawler’s architecture they exploit the libmetaverse[1] library to interface custom scripts to the Second Life client architecture. To collect the significant amount of data they employed a multi-avatar architecture, each with its own assigned task. Tasks included collecting current region listings, avatar location and activity, object metadata collection, and recording data used for statistical analysis. All recorded data was retrieved and stored immediately into a database for future analysis, and the crawler was ran on a 24 hour frequency for a little under a week, on multiple occasions. Recording at this frequency allowed the researchers to record change over time as well as confirm data collected in previous crawls.

Chen and Zhang show a similar approach to extracting information from virtual worlds [11]. Using libmetaverse, they were able to create bots that would observe avatar behavioral patterns. Their collection framework consisted of these Second Life bots and a traditional spider architecture which collected meta data from the observed avatars profiles. Using both sets of data they were able to show behavioral models of avatars of different gender, age, and other profile attributes.

One of the complications with attempting to mine data from virtual environments is the lack of an Application Programming Interface (API) provided by the creators of the world. Like Second Life, World of Warcraft does not provide an API that allows character statistics and world data to be collected programmatically. However, some creators of these worlds also publish websites that present some of this in-game information. Lewis and Wardrip-Fruin use

[1] http://www.openmetaverse.org

traditional HTML scraping techniques to collect World of Warcraft data, enabling a full analysis which would otherwise be impractical to perform [12].

2.2 Social Networks

As our work pertains to social networks, we need to first study web-based social networking sites such as YouTube or Facebook. In particular, we need to understand what social networking sites are, where they came from, and what implications arrive with their appearance to the Web.

Boyd and Ellison elaborate on the definition, history, and research prospects come with the evolution of social networking sites [13]. They define social networking sites as any web-based service that allows its users to generate public profiles. In addition, the users may list others to whom they wish to connect as well as traverse the list of connections they have created and the lists their friends have created. While most social networking sites exhibit these behaviors they vary greatly between each other in features and subscribers.

The history of social networking sites begins in 1997 with SixDegrees. SixDegress was considered the first social networking site and took many social features from various applications and consolidated them to one place. As SixDegrees failed because 'it was ahead of its time, LiveJournal found its success in 1999 as many other social networking sites did in the subsequent years. The age of social networking sites began in 2003 with the proliferation of social media and user-generated content. MySpace and Facebook suceeded due to innovative features such as customizable profiles and an innovative strategy in initially restricting invited subscribers creating a demand for the hot new site. While it was realized that social networking sites focused on people and not on content, a new research area evolved to answer important questions about these applications.

Studies of social networking sites began as inquiries into how impression management, self-presentation, and friendship performance affected user experience. Network analysis would evolve to answer questions about large-scale information such as entity connections and usage. Other relevant topics include bridging the gap between offline and online social networks and how much privacy would be given to users and how to restrict the flow of private information. This rapidly developing field of research was required to get working on analysis to answer questions about a large scale applications and guide how systems would be tailored to a variety of changes.

2.2.1 User Activity

Explicit friendship associations between two users link the entities in a conventional social network. A variation of these common social networks is interaction graphs in which the links resemble actions between two entities in the network. Wilson et al suggest that common social links may not be a strong enough resemblance of actual interaction between nodes in the network and interaction graphs may be the solution to this problem [14]. This study observes at interactions via the photo comment and wall post functionalities on the Facebook social network site. Exchanges across these two channels were recorded across more than ten million profiles and analyzed to determine the efficiency of the interaction graph.

This research yielded significantly different results than a traditional social network link definition. After using a cumulative distribution function on the recorded data points, the researchers found that a large majority of the interactions occur between only a small portion of traditional link types, hinting at a large number of inactive relationships. Additionally, many of the entities in the network only received photo comments from less than five percent of their friends in the original social network. The actual activity occurring over such a significantly small portion of the friendship links proves that an interaction graph can provide more up to date and reliable information regarding the relationship between network entities.

Benevenuto et al. present additional research regarding user activity and how it may provide insight to the behaviors of the individual entities on a large scale [15]. Using a system that aggregated HTTP requests from a number of online social networks, they were able to collect user activity data from approximately 37,000 users. Each action under each HTTP request is categorized and analyzed to provide information resembling the interests of the users. Additionally, as these requests have timestamps, a temporal mapping of user activity could be made and probabilistic models of individual activity categories could be made.

In this study, the researchers also consider the distances covered by the activities and attempt to see how local a user's actions are. Eighty percent of a user's activities in the various social networking sites occur within one hop of the user. Interestingly, most visits to immediate or non-immediate friends originate directly from the user's home page. As for information propagation, the approximately twenty percent of actions occurring over a distance of two hops suggests that propagation is quite high.

2.2.2 Other Link Definitions

Traditionally, social network studies conducted using social networking sites have assigned ties between entities in the network as explicit friendship declarations between users. As most of a network's analysis operates on the ties that bind entities together, it is important to explore and understand alternative link definitions. Lewis *et al* investigates alternative ties between entities in a study that included a near-complete social network extracted from a set of students from a university [16]. Using this data, complemented with information from the university regarding the students' housing situation, the researchers built a social network with links representing dorm mates, group mates, or roommates. They also used the friends lists of each individual as well as tagging information extracted from their photos on Facebook as relationship links.

The results of their report included a multitude of various demographic reports based on collected information but, more importantly, preferences of each individual and comparisons across each network type. They reported highest similarity between friends who appear in each other's photos and not between the users who were on each other's friends list. Contrary to expectations of a user's proximity to others to increase similarity between the two, this was not the case for the networks generated by the housing data. The study also reports that students' preferences conflict with the information portrayed on Facebook.

In a different study, Adamic and Adar [17] use information collected from a collection of home pages from Stanford and MIT and attempt to predict user relationships. Specifically these relationships are inferred from each individual website's in-links, out-link, mailing lists and text. The authors found that text presented on the home page is the worst measure of similarity whereas inward links perform the best. Similar to the previous study, Lewis *et al* built a network based on alternative link definitions that also displayed traditional social network properties [16]. They also found that as the distance between two entities increases, so does the similarity scores for their websites. In that network, friends almost always have friends in common and clustering and community information is available to be extracted.

2.3 Social Network Analysis

The research on the analysis of social networks is quite vast and it includes many approaches to extracting and modeling behaviors of the network. Butts' article provides an overview of common analysis methods [18]. An in-depth description of the nodes, links, and the structures and paths they create within the social network graph are described in detail. The author also discusses methods by which the networks may be represented in computer memory, including matrices or spare-graph representations, since social network analysis at large scales stresses the power of modern day technology. Many research projects take advantage of the metrics and formulas described in this paper to present the behaviors of their collected data.

In addition to analysis, Butts' paper suggests that one must properly evaluate their network to provide a base line for further study. Methods of generating random graphs while following certain parameters are shown to provide a good basis for comparison. In addition to complete network analysis, node level indices and attributes may be calculated to provide information that may shed light on the behaviors between individuals in the network. The papers that follow employ many of these evaluation techniques but as the research area of social network analysis grows the methods used to analyze them will mature further.

2.3.1 Online Social Networks

Social networks formed by groups of individuals have long been analyzed for the purposes of psychological and biological research. With the Information Age and the evolution of online social networking sites, this research has expanded to include the study of networks formed by Internet-based social sites such as Facebook or MySpace. In order to compare a virtual world social network to these Web-based social networks, it is momentous to understand the authoritative approaches to analyzing online social network data. Mislove et al. present social networks from a variety of online social networking sites: Orkut, LiveJournal, Flickr, and YouTube [19]. This data was collected from both APIs provided by the websites or collected using a technique called HTML scraping which parses the html and gathers the required data. To provide a context of their findings, they compare their analysis to another research of the

Web as a whole. Additionally, the structural properties of the strongly connected component, or the subset of nodes in the graph that each have a directed link to and from one another, are explored and presented.

Classic properties of social networks are analyzed and presented to the reader. Many metrics are based on analyzing the degree of the nodes in the networks, i.e., the number of links into and/or out of a node. The higher the degree, the more connected the node and, more intuitively, the more relationships the person represented by the node has with other persons on the site. The power law summarizes the distribution of the node degrees in a network by showing that the probability of a node having a certain degree k is proportional to $k^{-\gamma}$, with γ being the power-law coefficient. In general, the power law, when verified in many social networks, shows that the network follows the distribution pattern found in many networks found on the Web and in the real world. Misla et al. also found that the power law to holds on all of the sites in their study. They also found several characteristics of the social networks that seemed to be consistent across most of the different social network sites. First, they found that that a relatively few high degree nodes existed within the network and many clusters of low degree nodes were found throughout. Second, a characteristic core was found to hold the network together. Thus, in all social networks, there seems to be a core component of highly connected users and a large number of inactive participants.

Nazir et al. conducted a seminal investigation of the currently most influential social network, Facebook [20]. The study of Facebook and the social networks behind it are quite important, as it has been shown to be the fastest growing online social network. They tracked more than eight million of users, a large user group but just a small fraction of all Facebook users. To track their subjects, they created three applications using Facebook's developer toolkit that were deployed by the research team. Users of these applications had their actions tracked and analyzed. Unlike the previously discussed paper, the links between the nodes are not explicitly defined relations between the users but rather the actions performed between the users. Over 35 million interactions were recorded over the three applications; in addition, geographical data about each individual was collected. The temporal nature of the users' actions allowed the analysis of active users on a daily basis as well as more traditional exploration of community structure within the network.

As was found in other studies, a densely connected sub-graph, or core, was found in this online social network. The power law was shown to hold and other node degree distributions followed conventional patterns found within social networks. In addition, the geographical diversity of the communities within the network was studied, revealing a lack of correlation between community members and their physical locations within the same country. Importantly, the community size of the gaming application is shown to be biased when compared to the two social applications deployed due to the nature of user interaction. Many more communities of relatively smaller size are found in applications based on social interaction when compared to applications based on gaming interaction. Finally, the paper was able to discuss the nature of users and their interactions on an online social networking site as popular as Facebook.

2.3.2 Effect of Link Definition on Analysis

The choice of the definition of a link between nodes has a dramatic effect on the size and topography of the social network constructed for a given online site. Viswanath et al argue against using traditional social network link definition of an explicit friend link [21]. They point out that interactions among users make better links because they are able to differentiate between active and stale inter-user relationships. They built and studied an interaction-based social network for 60,000 Facebook users in the New Orleans area with over 800,000 interactions. They compared their interaction-based network to a traditional social network of 90,000 users with 3.5 million friend links. In addition, they recorded activity for the interaction-based network over three-year period, with ninety-day snapshots, providing a relatively long period to show temporal variance within the activity network.

In the comparison between the two networks, it is found that node degree in the activity network is significantly lower than the social network. This shows that users only actually interact with a fraction of their listed friends and suggests that friendship links may be a weaker edge definition than previously assumed. Another notable result of their research was that structural properties of the interaction graph stayed approximately the same over the time period. Although new interaction graphs were generated at each snapshot, the network topography did not vary much at all. Due to the nature of interaction graphs, the actions between users as the links between the nodes are considered a better representation of real interaction, and thus relation, between users. The researchers argue this point with the additional consideration that spikes in activity may be due to application features such as birthday notifications and should be examined.

An alternative approach to varying the link definition was given by Matsuo et al [22]. Instead of simply using an interaction graph, the researchers construct three social networks from the same set of authors and co-authors: web-crawled (*collaborator*), user-registered (*knows*), and physical proximity (*meets*). In addition to the graphs as a standalone network, they perform what is called a multiplex graph or a summation, average, maximum or multiplication of any two of the three networks. An analysis of these networks and combinations thereof were performed and reported. In a direct comparison of the three networks, they find the *knows* and *collaborator* networks share the highest number of similarly linked avatar pairs. After locating network authorities, it was shown that most authorities use *collaborator* links, middle- authoritative use *knows* the most and less authoritative use *meets* links the most. It is stated that as link definition has such a drastic effect on the properties of the networks that the context of the research should guide the decision on what to base node relationships on.

2.3.3 Sampling Techniques

Most social networks are extremely large. It is difficult or impossible, for researchers to get the complete data for the network in many cases as the developers of digital social networks have all of the data. Other parties can generally access only a subset that is made visible. Even if the data is available, many analysis algorithms are infeasible for data of that scale. Thus, social network analysis research on large sites is conducted using subsets of the users sampled in a variety of methods. Many researchers in the area of social network analysis tend to overlook the importance of the sampling techniques used when attempting to analyze these large networks. Leskovec and Faloutos attempt to provide an outline of many sampling techniques and highlight the benefits and disadvantages of using each one [23]. It is important to ask each of the following questions: what makes a sampling method good; should nodes or edges be sampled randomly or are there better strategies; what a good sample size is; and how to measure the goodness of a sampling technique. In this paper, the authors present two custom methods alongside many other algorithms found in published literature and use each technique on five different datasets. To measure the goodness of each sample, they use the Kolmogrov-Smirnov D-statistic to measure the agreement between the distributions of each sample. The algorithms are run ten times with varying starting parameters on the various datasets and the patterns are plotted and presented to the reader.

The results show that the Forest Fire, Random PageRank, and the Random Node sampling techniques provide the least amount of bias in the node degree distribution. Between the two algorithms the authors suggest, the Back-in-Time sampling algorithm performs best. This algorithm attempts to roll back the growth patterns of a graph with little or no temporal information to a desired sample size. Among the more traditional sampling methods, the random node selection accompanied with vicinity exploration provided the best samples among the tested algorithms. Additionally, it was found that a sample size of approximately fifteen percent of the original data generally provides enough information for proper analysis of the network.

In a similar study where multiple sampling techniques are compared side-by-side, Gjoka et al. set out to make a representative sample of Facebook Users [24]. The sampling techniques compared include Breadth First Search, Random Walk, Re-Weighted Random Walk (RWRW), and Metropolis-Hastings Random Walk (MHRW). The base truth the authors have used for comparison in this paper is a uniformly random sample of just under one million Facebook users. To gather the required information, the researchers use an HTML scraping method.

The authors found that the MHRW and RWRW provide near identical node degree and regional network distributions against the uniformly random network. The Breadth First Search and Random Walk sampling techniques show bias in almost all of the metrics related to node degree. Between MHRW and RWRW, there are various tradeoffs, but overall the authors recommend MHRW for its versatility.

3. APPROACH

3.1 Crawler for a Virtual World

To collect data for this study, we have chosen to utilize an intelligent virtual world crawler we developed for a previous project [2, 25]. Our crawler mimics the behavior of traditional web crawlers in that it collects objects from its current location and adds links from the current location to a queue of locations to be collected. Once all information from the current location has been collected, the crawler moves to one of the links stored in the queue. Our crawler is able to move semi-autonomously, either following a set of waypoints provided to them or by following other avatars in the three dimensional environment.

Using this crawler we were able to populate a relational database with data from the virtual world. Essentially anything that is sent from the Second Life servers to the client in a traditional user session is collected as the crawler moves throughout the virtual world. Any object loaded within a certain proximity to the crawler's position has its metadata collected including scripts, graphical information, and any objects attached to it. Regional information is collected, as well as linked landmarks and local chat messages between avatars and objects. More importantly, avatar and group information is collected through a Breadth-First-Search approach. Groups can be queried from the Second Life server, and any avatar that the crawler comes in contact with may publicly list his or her group subscriptions. The group subscriptions listed by a crawled avatar are polled from the server and return an avatar list which then may have a public group listing of its own. This process repeated until no new groups are collected results in an avatar-group social network that we have built this research upon.

Using this snapshot of data collected from the virtual world, we were able to use popular network analysis techniques to model the behavior of social entities within Second Life. The analysis of such a large dataset presented many scalability challenges. To analyze our large network, we tested several software libraries and conducted an in-depth study of which of the various metrics needed to elucidate the behaviors of the virtual world social network as a whole could feasibly be calculated. The data set was sizeable enough that it demanded a representative subset of the entire network be taken and used in the analysis process. We reviewed common sampling techniques to create a tractable-sized sample that still exhibited the characteristic properties of our network to keep the analysis results meaningful.

3.1.1 Architecture

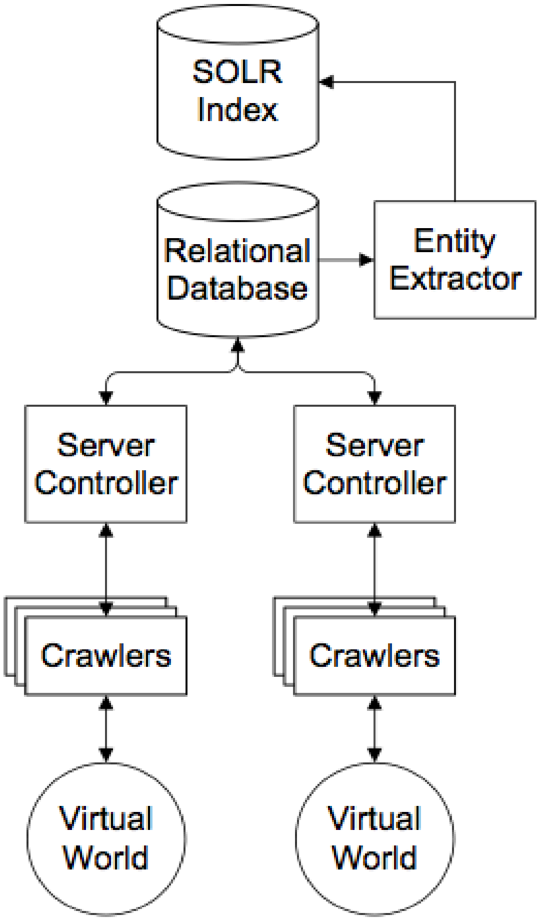


Fig. 2. Crawler Architecture [28]

Figure 2 diagrams the architecture for the data collection system for virtual worlds developed by Eno [28]. A collection of avatars controlled by custom scripts interacts with the virtual world servers. These scripts define individual avatar navigation and movement within the virtual world's regions. Behind the system of avatars, one or more server controllers monitor and govern where, how, and when the avatars are dispatched to crawl a region. The data collected by these avatars is stored in a MySQL relational database that supports efficient data storage and access. In this research, the data used for the social network is queried directly from this database and processed through custom software to discover metrics relevant to social network analysis.

3.1.2 Movement and Navigation

Movement of the individual avatars is controlled by an API that allows for a range of movement functionalities similar to that given by the graphical client to the end user. The movement object collects a series of waypoints sent by the navigation object and is left with the responsibility of path finding and obstacle avoidance. Three threads work in parallel; one thread each for movement in the horizontal and vertical dimensions and one thread to determine the success of the avatars movement in both dimensions. This final thread also has the responsibility of detecting whether or not the avatar has become stuck within the virtual world. The movement of the avatar can become blocked either from landscape objects or user created content so it is important to have routines in place to either route around the obstruction or teleport to the nearest destination.

Working in conjunction with the movement object is the navigation object that is directly responsible for the large-scale waypoint generation for the avatar. This object attempts to give the crawler waypoints in a manner that will cover the most volume of three-dimensional space for an individual region, spiraling out from the center of the region and varying the altitude. The navigation object may respond the new information that the crawler obtains and issue a new route of waypoints while accounting for obstacle avoidance to prevent the crawler from becoming stuck. The movement object simply requests new waypoints so it need not be aware of changes to the paths by the navigation object. The navigation system may also send a new set of waypoints should the movement object's path become obstructed and require a reroute to navigate correctly around the obstacle.

3.1.3 Server Controllers

The server controller manages the main configuration, startup, and monitoring of the crawling process. This includes processing runtime settings from configuration files and loading data storage information. The controller also obtains from these files the login information for each individual crawler and any to-crawl region listings available. This list of regions feeds into a global virtual world region list that is maintained by the server controller and dictates the crawler's teleportation destinations as regions are successfully crawled and created within the virtual world. In addition to maintaining this region list, the server controller is responsible for stopping, starting, and scheduling of the crawling processes. The memory usage of each thread spawned for an individual crawler is also monitored by the server controller, which allows for flexibility and sustainability of the crawling processes.

3.1.4 Storage

The storage aspect of the crawling system consists of two entities: the storage system and the queue manager. The storage system uses a single MySQL object along with a system of queues. These queues allow for the crawling avatars to not be limited by the read and writing times of the database. Additionally, the server controller may slow or stop the crawling processes as required based on the saturation of the queues. Global entity information (e.g. avatars, groups) may be persisted through crawls by writing the relative queue to disk on shutdown and reloading into memory on startup.

The queue manager maintains the queues and has two main functionalities. First, it maintains a list of global entities that have been discovered either from group listings or crawler discovery but have not been requested from the virtual world servers. In general, the global entities are discovered more rapidly than can be processed by the storage system. The queue manager may then distribute the request of individual global entities to various crawlers, inherently increases crawl speed and distributing load across the entire crawling system. Second, the queue manager is responsible for duplication checks. A hash of the avatar profiles, avatar names, groups, and parcels linked from avatar picks, or favorites, is stored in memory. When a crawler sends a piece of data to the queue manager to be stored, the queue manager checks the hash of the data against this list, and may prevent adding it to the database if it already exists within this list. If the item has not been updated within a refresh threshold, the object will be updated within the database.

3.2 Social Network Analysis

In this section, we describe the network analysis metrics we will use to characterize our virtual world social network and compare our experimental observations to prior work in social network analysis. We also introduce our approach using third party libraries and our experience with them, as well as how we developed custom software to achieve the calculation of many of the metrics.

3.2.1 Methods Selected

We chose to analyze the avatar social network using metrics used in similar studies involving social networks discovered on the web. These metrics range from basic to complex. Here we provide a brief explanation of the key metrics used in this study.

- *Degree:* The degree of a node in an undirected graph is the number of adjacent edges. Since edges represent the number of groups shared by two avatars, the degree of a node is equal to the number of relationships for each avatar.
- *Fraction of Avatars:* The number of avatars in the sample used for analysis compared to the total number of avatars in the virtual world.
- *Average Node Degree*: The average of all of the avatar's degrees in the sample.
- *Number of Groups:* The number of unique Second Life groups that are represented within the sample used for analysis.
- *Average Groups Per User:* The average number of groups per user within a sample. Groups exceeding a predefined size limit are not considered in this average.
- *Average Path Length:* This is the average of the shortest paths between all pairs of nodes in the network. Due to the size and complexity of our networks, we took a random sample of one hundred nodes and calculated the shortest paths from these random nodes to all other nodes to estimate the average path length.
- *Radius:* The radius of a graph is defined as the minimum eccentricity of the graph. Given a node n, the eccentricity of n is the maximum shortest path between n and any other node [26]. For our networks, we calculated the radius of the largest graph component. We also took one hundred random seed nodes and calculated the

eccentricity of each of them. The minimum eccentricity of these seed nodes gave us our radius measurement for our social network.

- *Diameter:* While not simply twice the radius, the diameter metric does find the shortest path between the edges of the network. The diameter is the maximum of the eccentricities across all nodes. Similar to the radius, we take a random sample of one hundred seed nodes and calculated the eccentricities between a seed node and all other nodes in the network.
- *Joint Degree Distributions:* The joint degree distribution displays the rate at which nodes of a certain degree connect to nodes of another degree. This degree distribution is represented as 2D matrix of the node degrees, the coordinate value, or shade of gray at a pixel, represents the strength of connection between nodes of that degree.
- *K-Nearest Neighbor Distributions:* We define kNN (k-Nearest Neighbors) degree distribution to represent the tendency of nodes of a certain degree to connect to higher or lower degree nodes. A value of a point on the scatter plot attributed with a given degree is the average degrees of all nodes connected to a node of this degree.

3.2.2 Implementation of Analysis

To begin the analysis, we first had to determine the definitions of the edges connecting the entities in the social network. As explicit friend relationships between avatars were not publicly available, and therefore not available to be collected by the crawler, we had to choose between defining edges as pairs of avatars that belonged to the same group or avatars seen in the same area at the same time. The latter contained extremely sparse data so we chose the shared group definition. This choice allowed us to maintain a meaningful representation of the entities of the virtual world and their relationships to one another. Our network consists of 1,628,532 unique avatars that we crawled from the Second Life server. To add edges to the graph, we processed the list of all 310,702 Second Life groups, and for each group we incremented the weight of the edges between all possible pairs of avatars in the group. The addition of larger group sizes leads to an explosion in the number of edges added to the graph. For example, a group with 100 avatars will produce 4,950 edges, while a group with 1000 avatars will yield 499,500 edges. Our largest group of 28,328 avatars would produce a staggering 401,223,628 edges. Unfortunately, this exponential increase in edge counts would lead to problems involving hardware limitations and analysis algorithm runtime.

Initially pulling the avatar group listings from the database into custom Java software, it became apparent that memory limits would become an issue in this research. Using dual hash tables for avatar to group ID relations, and vice versa, the memory required grew quickly. It is necessary to store the entire network in RAM so the analysis algorithms may be run at a rapid pace, and once this has overflowed into virtual memory on the hard disk it difficult to tell when the complex algorithm will complete. Even converting the UUID, a unique identification string, of the groups and avatars into unique short primitive data types with 16-bits per object, the memory limits were still being surpassed due to the sparseness of the graph and the overwhelming number of edges. It was decided that others in the network analysis community might have a better approach to storing and analyzing large social networks. Additionally, membership in a group that contains so many members is not a very strong indication of a link between two avatars.

As many libraries for network analysis could only support graphs of a significantly smaller size than our network, the search was narrowed down to a few candidates, SNAP[1], Neo4j[2] and iGraph[3]. SNAP is a network analysis library published by Stanford and written in C++. The library is able to scale to networks of millions of nodes and billions of edges. This sounded like an ideal solution given our extremely large network size but with further investigation it was found that many of the targeted analysis metrics were not included in the library. Neo4j appeared promising, as it is a NOSQL graph database, meaning that the database is stored on hard disk as a directed graph itself rather than the traditional structured file approaches. This allowed for the support of all of our network data on disk, however the analysis algorithms efficiency was adversely impacted by the read and write speed of the hard drive. Finally, the open source iGraph library for Java developed by Csárdi and Nepusz was designed initially for representing large networks of particles on the atomic scale to model behaviors and patterns of the nodes effects on one another. This library supports a node count limited by memory alone, 8 bytes per vertex, which supported the count of nodes our social network in the virtual world given our available memory. The data could easily be imported into the iGraph data model by using a Large Graph Layout, or LGL, file that listed the nodes and edges between them in a minimalistic flat file. iGraph seemed extremely promising with its plethora of analysis functions and its ability to support the network in memory.

Although iGraph supported enough nodes to support our virtual world social network, it still had issues with the number of edges between the entities. The complete network could be stored in memory but when some of the algorithms were run, the program would consume all

[1] http://snap.standford.edu
[2] http://www.neo4j.org
[3] http://igraph.sourceforge.net

memory available. Many of the complex analysis methods provided by the library such as average path length and radius did not support networks of such a large size. In addition, the unpredictability of the Java garbage collector and low control of memory allocations when using the Java language added additional complications to the completion of such algorithms. It became apparent that a different approach that allowed for lower level memory control was necessary.

3.2.3 Software Developed

Since the packages described above did not scale sufficiently, we developed our own graph analysis software. We chose to use the C++ programming language because it provides mechanisms to control data allocations of primitive data types representing the individual avatars and their connections to one another. Thus, we could keep the memory used to the minimum necessary for a given operation. Additionally, we would be able to rule out the unpredictability of the Java garbage collector in the run times of our algorithms and have the flexibility of custom logging functions.

Specifically, we devised an algorithm to take advantage of the speedy lookup of a binary tree for link updates and inserts. Within memory, we define an array of binary trees, one for each avatar, that represents the connections to all of its neighbors within the network. Each link between a pair of avatars holds an integer value representing the number of group subscriptions both avatars have in common. As group listings are brought in from the LGL file, they are inserted into the tree and incremented if the link already exists. Anytime there is a lookup of a link between avatars we always reference the binary tree array by the avatar with the lowest ID, and search for the larger ID within that tree. This allows us to maintain a lookup time complexity of O(log n) and keep an organized data structure representing the social network.

Including all of the edges from all Second Life groups would exceed even our custom graph software capacity, so we had to represent the social network with a subset of edges. The added control for tweaking algorithms and data structures in our custom software proved enough reason to still use it over the graphing libraries. Our target algorithms also were not included in a single package, and generally the development support and documentation was lacking when using these open source software packages. Since random edge sampling has been shown to perform poorly [1], we chose to exclude all groups above a specified maximum size when constructing the social network graph. If we assume that the strength of avatar relationships is inversely related to the group size, excluding large groups will have the effect of ignoring weak edges between avatars when creating the social network graph. Once the sample was loaded into memory, we were able to run the variety of network analysis algorithms, listed in section 3.2.1, to retrieve meaningful statistics from the sampled network.

4. EVALUATION

4.1 Data Collection

Using our Second Life crawler [2], we simulated the movements of an avatar and collected and stored all information that was sent from the Second Life server to the graphical client. A summary of the data we collected over several months is shown in Table 1. Although this is not as complete as the original data stored on Second Life servers, we feel this data collection provides large representative cross section of regions/avatars/groups that exist within Second Life. The next section will present in detail how we extract a representative sample of this data to build a social network of avatars within this virtual world.

Table 1 - Summary of Collected Data

Total Regions Crawled	20,901
Total Unique Avatars Crawled	1,628,532
Total Unique Groups	310,702
Total Group Subscriptions	9,534,064

4.2 Experiments

A consideration when setting a maximum group size limit to be included within the social network is the real relationship between avatars in the same group. When groups of large size are added to the social network, it increases the strength of existing links between avatars and adds links between several pairs of avatars. In order to keep this fact in consideration, we have taken various samples of our social network for comparison between them and those found in other studies. Traditionally, samples would be taken at a 25%, 50%, 75%, and 100% but due to the power law distribution of the group sizes [27] we believe it is more useful to take samples of the 70th, 80th, and 90th percentiles while including a 99th percentile sample as our maximum. This decision was made based on the significantly small changes between basic measures of the generated networks. Figure 3 represents the percentages of groups included in a sample given a maximum group size.

Table 2 shows that ninety percent of groups have a membership of 41 members or less. This suggests a behavior of group members wanting to keep groups within a manageable size or being unable to find membership for their group. Larger group sizes are most likely public groups or large popular companies and the power law behavior is not surprising. As the group size increases, the amount of edges added to the graph is exponential. Additionally, this affects computation time of most of the metrics as many rely on finding all pairs shortest past for a given node. Given these facts, we have chosen to specify a maximum group size of 457 in order to represent 99% of the groups included in the crawled data while avoiding high edge groups in the top one percent.

Table 2 - Values of the Group Distribution and the corresponding percentiles

Group size	2	5	10	13	18	41	457	28328
Percentile	25%	50%	70%	75%	80%	90%	99%	100%

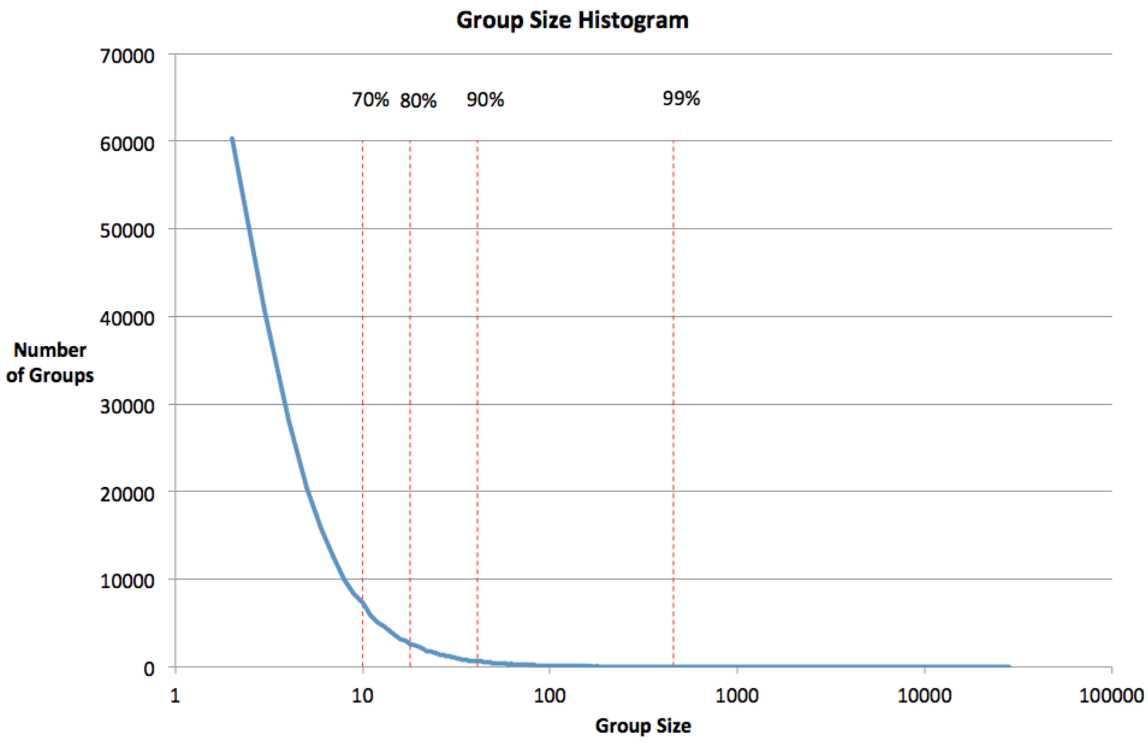


Fig. 3. Distribution of virtual world group sizes with different main percentiles marked

4.3 Evaluation of Social Networks

Using the data collected above, we have defined two social networks of avatars. The first social network (*All_Edges*) uses all edges between avatars. The second social network uses all edges with edge weights greater than one (*All_Except_1_Edge*). The resulting graph is a network of avatars that share two or more groups in common, which is a stronger connection than avatars that share only one group in common.

	Virtual World Social Network								**Online Social Networks [19]**		
	All_Edges				***All_Except_1_Edge***						
Percentile of Groups	**70th**	**80th**	**90th**	**99th**	**70th**	**80th**	**90th**	**99th**	**YouTube**	**Live Journal**	**Orkut**
Number of Avatars/Users	447,825	576,991	760,445	1,273,002	97,396	153,369	239,006	604,861	1,157,827	5,284,457	3,072,441
Fraction of Users	27.4%	35.4%	46.6%	78.1%	6.0%	9.4%	14.7%	37.1%	Unknown	95.4%	11.3%
Number of Links	1,819,062	4,582,699	14,975,625	315,455,515	109,732	250,017	647,953	10,070,845	4,945,382	77,402,652	223,534,301
Avg. Node Degree	8.12	15.88	39.39	495.61	2.25	3.26	5.42	33.30	4.29	16.97	106.1
Number of Groups	21,9058	251,102	279,778	307,602	21,9058	251,102	279,778	307,602	30,087	7,489,073	8,730,859
Avg. Groups Per User	1.96	2.30	2.78	4.28	1.96	2.30	2.78	4.28	0.25	21.25	106.44

Table 3 - Comparison of basic measures of social networks

In Table 3 we show a side-by-side analysis of our two social networks and their samples based on percentile. These results are compared with statistics observed in online social networking reported by Mislove [19]. The YouTube and Live Journal networks are directed networks. The edges in these networks have a direction from one node to the other and involve several other analysis techniques. The Orkut social network is an undirected network that also contains a large amount of edges. Our network has approximately half the number of nodes while the number of edges is still larger than other networks. This suggests that the removal of edges with a weight of one is justified in constructing a social network based on implicit link definitions.

More complex statistics of our network are presented in Table 4 alongside online social networking site data [19]. The average path length of the virtual world social networks does not deviate significantly from online social networks either. Our social network composed of edges of all weight (*All_Edges*) has a slightly lower average path length due to the amount of edges. The virtual world social network composed of edges with a weight greater than one (*All_Except_1_Edge*) strengthens the relation represented by a link, but adversely increases the average path it would take to get from one node to the other. Similar cases are found when comparing the radius and diameter of the graphs found in virtual worlds and those found on the web including Flickr, Live Journal, and YouTube. It suggests that even though edges of weight one can be removed, the furthest an avatar is from another in the largest component of our virtual world social network is a shorter distance than those found in online social networks.

Table 4 - Structural properties of the social networks

	VW Social Network		**Online Social Networks [19]**			
Graph	***All_Edges (99%)***	***All_Except_1_ Edge (99%)***	**Flickr**	**Live Journal**	**YouTube**	**Orkut**
Avg Path Length	2.96	4.89	5.67	5.88	5.10	4.25
Radius	7	11	13	12	13	6
Diameter	9	14	27	20	21	9

In Table 5 we compare the social network found in our virtual world to two studies involving Facebook applications and a variety of other online social sites [14]. The 99th percentiles of both versions of our network are presented and have similar statistics of the online social networks. The number of components is quite large in the network with edges of a weight of two or more. This is expected, as the removal of edges with a weight of one will disconnect the graph substantially.

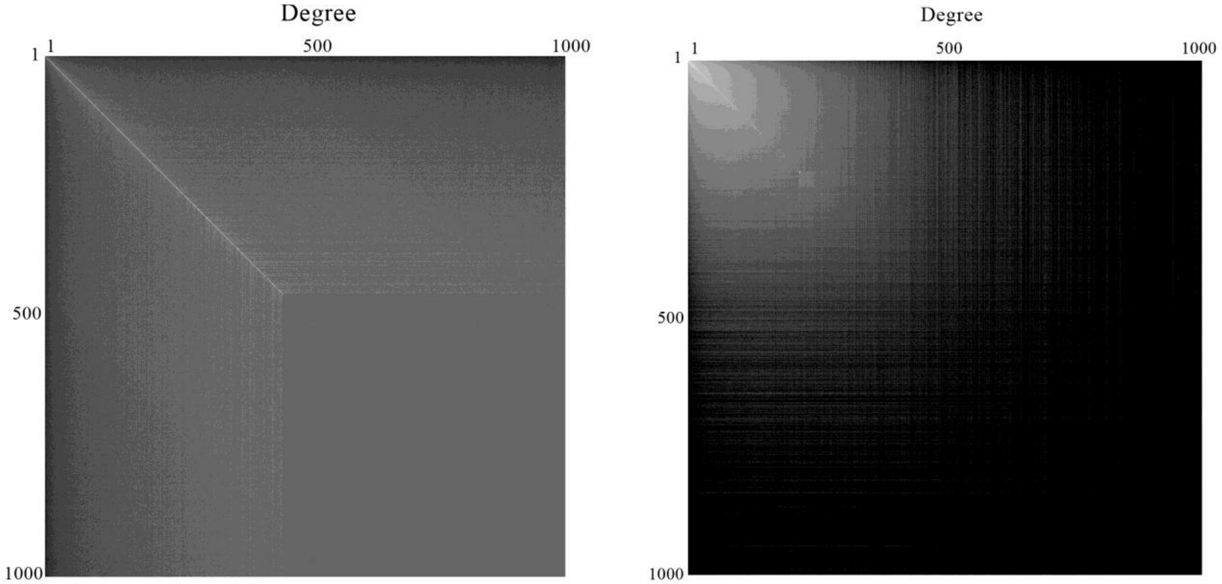
Table 5 - Component comparison of the social networks

	VW Social Network		**Online Social Networks [20]**		
Network	***All_Edges (99%)***	***All_Except_1_ Edge (99%)***	**Fighter's Club**	**Got Love**	**Hugged**
Number of Components	593	6042	29	13461	4018
Largest Component %	99.87%	98.73%	91%	92.10%	86.70%

4.4 Joint Degree Distribution

Figure 4(a) shows the joint degree distribution of the network of all edges included (*All_Edges)*. A clear diagonal is very distinctive and shows a clear connection between nodes of similar degree. High degree nodes connect to high degree nodes, and low degree nodes connect strongly to other low degree nodes. Another fact about the network that can be discerned from Figure 4(a) is the somewhat substantial cutoff of the diagonal and the relatively sparse bottom right corner of the image. This is due to the fact that for presentation we have removed degrees of one thousand or higher from the image but the maximum group size used for the 99th percentile is 457.

In comparison the network including edges of all weights, Figure 4(b) shows the joint degree distribution gray map of the same network with edges of weight one removed (*All_Except_1_Edge*). Significant differences include the absence of the substantial diagonal through the higher node degrees. This suggests that many of the nodes degrees consist of edges with a weight of one. It makes sense that as only a small portion of edges in the full edge network exist in this representation, but really shows that the high degree avatars are connected to many other avatars simply because they belong to one or a few relatively large groups.

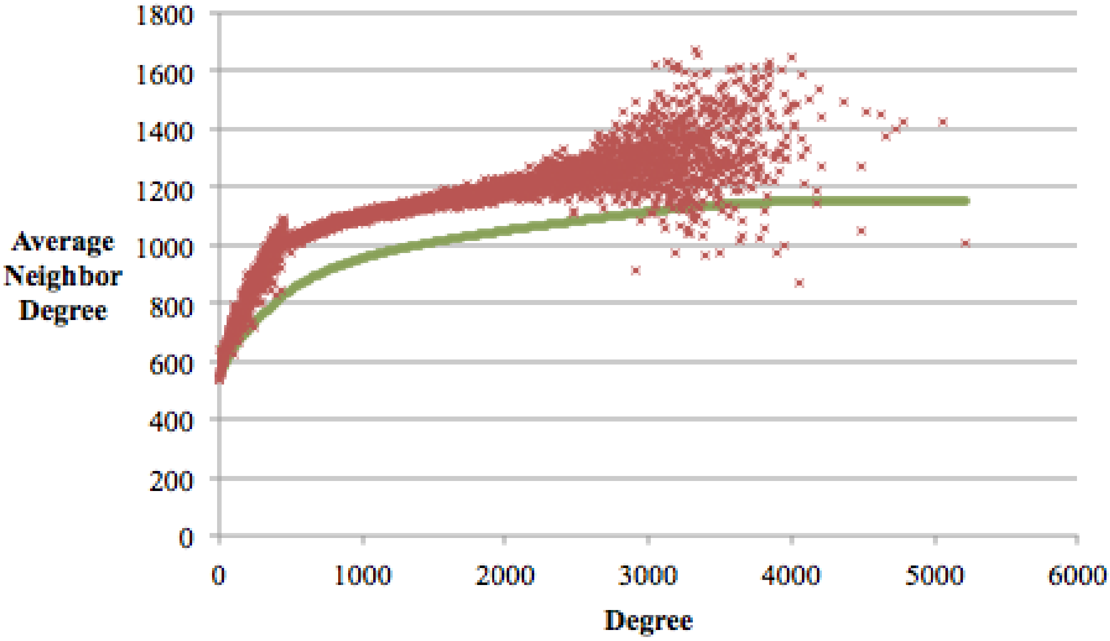


(a) (b)

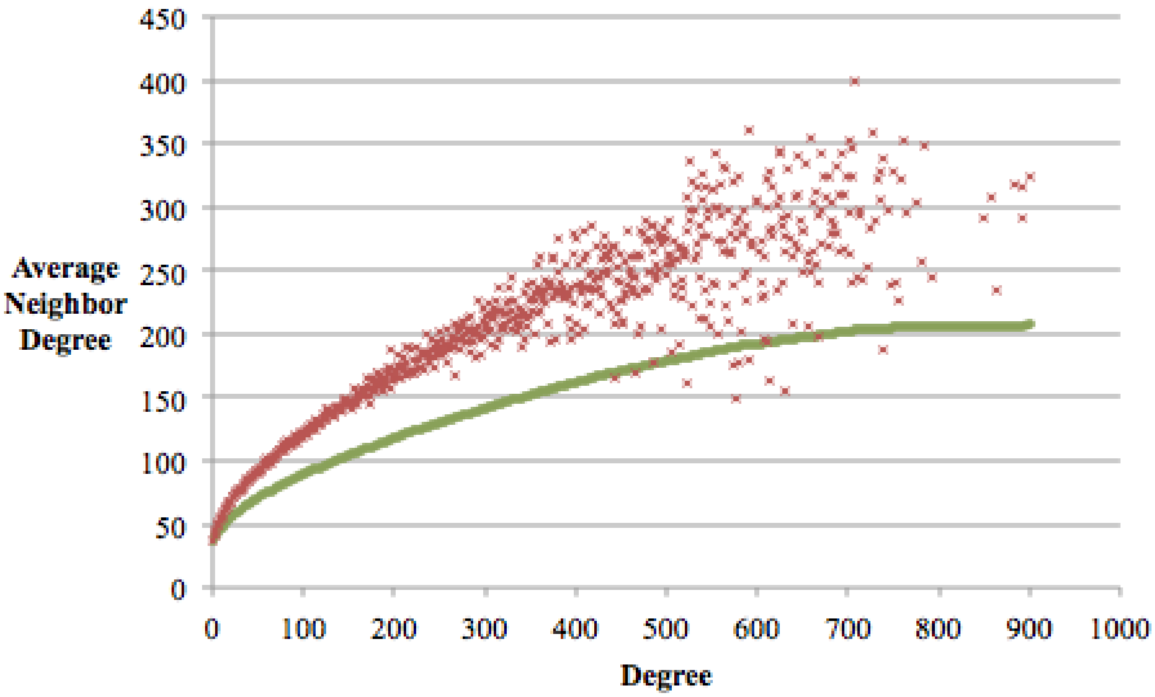
Fig. 4 - Joint degree distribution

4.5 k-Nearest Neighbor Distributions

In Figure 5(a) and 5(b) we have the kNN scatter plot of the virtual world social network with all edges included (*All_Edges*) and the virtual world social network with all edges with a weight greater than one (*All_Except_1_Edge*). The kNN value of an undirected network is the mapping of the degree of a node and the average degree of all nodes with the initial degree's neighbors. The green line in the figure below represents the average kNN for all degrees less than the current value. The behavior of this line is a good indicator of the behavior of higher degree nodes. If the average increases then so does the propensity of higher degree nodes in the social network to connect to other higher degree nodes. The axis scales vary quite a bit between the two networks because the removal of edges of weight one decreases the degree of all nodes significantly due to over 96% of the edges being removed.



(a)



(b)

Fig. 5 - Scatter plot of kNN and average

4.6 Discussion

In social network analysis, it is important to understand how the entities of the network are related to one another via links or shared groups. From Table 3, we can see the number of avatars and links when we consider all edges between avatars (*All_Edges*) are very large. These numbers decrease significantly at all main percentile values if we take into account only edges that have weight greater than one (*All_Except_1_Edge*). Hence, there are a large number of avatars sharing only one common group with other avatars.

The more complex statistics show that the social networks of *All_Edges* and *All_Except_1_Edge* are reasonably close to those measured from other online social networks. The component analysis shows that while several isolated clusters of avatars exist, there exists one component containing over ninety percent of avatars in the network. The same fact is found in other online social networks. Numerous clusters of nodes exist in the social network but most of the nodes in the network are contained within the largest component.

The k-Nearest Neighbor scatter plot of the two social networks (*All_Edge* and *All_Except_1_Edge*) presented in the Figure 5 provides additional insight to the connectivity of nodes with similar degree. As both averages increase, it shows that avatars tend to be related more with avatars that have a higher degree, or higher amount of avatars in which they share groups in common with. The sparseness of the lower right corner of the Figure 4(b) suggests that the degrees of nodes are significantly lowered when edges with a weight of one are removed. The diagonal of similar degree nodes connecting to each other still persists through this network as in *All_Edges* network as seen in Figure 4(a). This indicates that the removal of all edges with a weight of one does not necessarily affect the behavior of nodes tendency to connect to nodes of similar degree. This finding agrees with that found in the kNN distributions.

The social network found underlying Orkut is most similar to our network including all edges. As Orkut is the only undirected online social network of the four, it suggests that perhaps undirected online social networks are naturally more compact and closer together than directed networks.

5. CONCLUSION

Understanding how avatars in virtual worlds behave when they connect to these worlds is important in terms of characterizing avatars behaviors and interests. The goal of this research study is to build and evaluate a social network of avatars that were crawled from Second Life virtual worlds. We have made a comparative study of important graph measures that can give a view about how social networks are differently formed in virtual worlds and in online networking.

The analysis of the virtual world social network compared to social networks found in online social networking sites has elucidated a few facts. Although the friendship data between avatars was unavailable to be crawled, the link definition of groups in common between pairs of avatars is a sound approach as the networks generated were not extremely different in their basic structural properties. Additionally, the fact that a user is using an avatar to interact with a 3D environment does not necessarily affect their social behaviors.

While this research outlines some of the basic structural properties and metrics to evaluate a virtual world social network derived from Second Life, there are several metrics we would like to explore (i.e., the clustering coefficient, associativity, and scale free metric) and use to further compare our virtual world social network with others. We also hope to analyze how the link definition affects the connectivity of the avatars in the social network.

REFERENCES

[1] B. Mennecke, D. McNeill, M. Ganis, E. M. Roche, D. A. Bray, B. Konsynski, A. M. Townsend, J. Lester, "Second Life and Other Virtual Worlds: A Roadmap for Research," *International Conference on Information Systems,* Montréal, Quebec, Canada, December 9-12, 2007.

[2] J. Eno, G. Stafford, S. Gauch, C. Thompson, "Hybrid User Preference Models for a Virtual World," *User Modeling, Adaptation, and Personalization*, Girona, Spain, July 11-15, 2011. pp. 87-98.

[3] W. S. Bainbridge, "The Scientific Research Potential of Virtual Worlds," *Science Magazine,* July 2007. pp. 472-476.

[4] C. Collins, "Looking to the Future: Higher Education in the Metaverse," *EDUCAUSE Review,* September 2008. pp 50-63.

[5] N. Ducheneaut, M. H. Wen, N. Yee, G. Wadley, "Body and Mind: A Study of Avatar Personalization in Three Virtual Worlds," *Proceedings of the SIGCHI Conference on Human Factors in Computer Systems*, New York, New York, USA, 2009. pp. 1151-1160.

[6] J. Doodson, "The Relationship and Differences Between Physical and Virtual World Personality," University of Exeter, 2009.

[7] C. Thompson, “Virtual World Architectures,” *Internet Computing, IEEE* 15.5. pp. 11-14.

[8] S. Benford, C. Greenhalgh, T. Rodden, J. Pycock, “Collaborative Virtual Environments,” *Communications of the ACM* 44.7, 2001. pp. 79-85.

[9] R. Kazman, “Making WAVES: On the Design of Architectures for Low-End Distributed Virtual Environments,” *Virtual Reality Annual International Symposium, IEEE,* 1993.

[10] M. Varbello, F. Picconi, C. Diot, E. Biersack, “Is There Life in Second Life?” *Proceedings of the 2008 ACM CoNEXT Conference,* New York, New York, USA, 2008. pp. 1-12.

[11] H. Chen, Y. Zhang, “AI, Virtual Worlds, and Massively Multiplayer Online Games,” *Communities 1, 3*, 2011.

[12] C. Lewis, N. Wardrip-Fruin, “Mining Game Statistics from Web Services: a World of Warcraft Armory Case Study,” *Proceedings of the Fifth International Conference on the Foundations of Digital Games, ACM,* 2010. pp. 100-107.

[13] D. M. Boyd, N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication,* October 2007. pp. 210-230.

[14] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, B. Y. Zhao, "User Interactions in Social Networks and Their Implications," *Proceedings of the 4th ACM European Conference on Computer Systems,* New York, New York, USA, 2009. pp. 205-218.

[15] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, "Characterizing User Behavior in Online Social Networks," *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference,* New York, New York, USA, 2009. pp. 49-62.

[16] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, N. Christakis, "Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com," *Social Networks*, October 2008. pp. 330-342.

[17] L. A. Adamic, E. Adar, "Friends and Neighbors on the Web," *Social Networks,* July 2003. pp.211-230.

[18] C.T. Butts, "Social Network Analysis: A Methodological Introduction," *Asian Journal of Social Psychology,* February, 2008. pp.13-41.

[19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattachariee, "Measurement and Analysis of Online Social Networks," *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, New York, New York, USA, 2007. pp. 29-42.

[20] A. Nazir, S. Raza, C. N. Chuah, "Unveiling Facebook: A Measurement Study of Social Network Based Applications," *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, New York, New York, USA, 2008. pp. 43-56.

[21] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, "On the Evolution of User Interaction in Facebook," *Proceedings of the 2nd ACM Workshop on Online Social Networks,* New York, New York, USA, 2009. pp. 37-42.

[22] Y. Matsuo, M. Hamasaki, H. Takeda, J. Mori, D. Bollegara, Y. Nakamura, T. Nishimura, K. Hasida, M. Ishizuka, "Spinning Multiple Social Networks for Semantic Web," *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006. p. 1381.

[23] J. Leskovec, C. Faloutos, "Sampling from Large Graphs," *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, New York, USA. 2006. pp. 631-636.

[24] M. Gjoka, M. Kuran, C. T. Butts, A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Samplings of OSNs," *INFOCOM, 2010 Proceedings IEEE*, San Diego, California, USA, May 2010. pp. 1-9.

[25] J. Eno, S. Gauch, C. Thompson, "Intelligent Crawling in Virtual Worlds," *International Joint Conference on Web Intelligence and Intelligent Agent Technologies*, 2009. pp. 555-558.

[26] P. Hage, F. Harary, "Eccentricity and Centrality in Networks," *Social Networks*, January 1995. pp 57-63.

[27] A. Clauset, C. R. Shalizi, M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, 2009. pp. 661-703.

[28] J. Eno, "An Intelligent Crawler for a Virtual World," *Doctoral Dissertation, University of Arkansas*, December 2010.