The Social Cloud for Public eResearch

by

Koshy John

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Engineering
in Network Engineering.

Victoria University of Wellington
2012

Abstract

Scientific researchers faced with extremely large computations or the requirement of storing vast quantities of data have come to rely on distributed computational models like grid and cloud computing. However, distributed computation is typically complex and expensive. The Social Cloud for Public eResearch aims to provide researchers with a platform to exploit social networks to reach out to users who would otherwise be unlikely to donate computational time for scientific and other research oriented projects. This thesis explores the motivations of users to contribute computational time and examines the various ways these motivations can be catered to through established social networks. We specifically look at integrating Facebook and BOINC, and discuss the architecture of the functional system and the novel social engineering algorithms that power it.

Acknowledgments

I would first like to thank my parents, John Koshy and Susan John, for their unwavering love and support in all my endeavours.

I would like to thank my supervisor, Kris Bubendorfer, for his valuable guidance and support throughout my thesis. Kyle Chard and Ben Palmer have my thanks for their contributions and feedback in the course of authoring the IEEE e-Science paper on the Social Cloud for Public eResearch.

Many thanks are also due to Andy Linton for his help with managing the development and test server for the Social Cloud for Public eResearch.

Finally, I would like to thank my friends in the Network Engineering Research Group for their feedback throughout the course of my research.

Contents

Chapter 1

Introduction

Scientific research increasingly relies on complex computation and large scale storage of scientific data, the scale of which cannot be provided by individual personal computers or even small clusters. Distributed computing models based on clusters, grids and more recently clouds, provide large scale capacity to scientists. However, obtaining funding to support IT infrastructure is often difficult due to the current structure of the funding agencies and licensing and ownership requirements imposed by commercial organizations. Access to national Grid infrastructures (TeraGrid, OSG) only supports selected projects and imposes strict time/resource restrictions on them. Several studies have also shown that conducting scientific research on commercial clouds often costs more than purchasing local resources [1, 2] and funding agencies are only now exploring models by which researchers can access public cloud time. This combination of factors significantly limits the processing power available to researchers.

Volunteer computing [3] is an alternative means of obtaining large computing resources – by getting the public to support specific projects by donating their spare computational and storage resources. The amount of computational time available to researchers is a function of the number of volunteers contributing at any given point of time. While there are a sizable number of volunteers who participate in volunteer computing (e.g

2.2 million BOINC participants [4]), this is insignificant when compared to the 800 million active Facebook users [5].

The primary goal of the research presented in this paper is to integrate Social Networking and Volunteer Computing and thereby bring eResearch to the masses. The way in which we do this is based upon our earlier work creating the Social Cloud [6]:

A Social Cloud is a resource and service sharing framework utilizing relationships and policies established between members of a social network.

We have named this fusion of social networking, social cloud and volunteer computing, the *Social Cloud for Public eResearch*. The potential for growth is significant, an uptake of only 0.5% of Facebook users would equal the entire existing BOINC user base. In addition, BOINC has no infrastructure by which new projects can be advertised, for example, of the 2.2 million BOINC users, over 1.1 million contribute to SETI@Home, while some newer projects have as few as 786 volunteers [4]. The public visibility of a project has a clear impact on the number of volunteers that it garners. We see this as another intrinsic advantage of adopting a social network like Facebook, where posts, news feeds, and social incentives can be used to bring a new project into the public eye.

For volunteer computing to be revitalized there needs to be an amalgamation of a mature volunteer computing platform and a high-potential source of new volunteers. In designing a proof of concept for the Social Cloud for Public eResearch, we chose BOINC as our volunteer computing platform. This decision was based on the maturity, modularity, and proven scalability of the middleware.

1.1 Thesis

The goal of the research is to study ways in which social networks can be integrated with volunteer computing middleware to raise the computational resources available to researchers. Social networks have a lot of potential to act as a means to reach out to the common man and have them participate in scientific efforts in their own small way. The public good possible through successfully tapping into a social networking service like Facebook to bring more people into volunteer computing is immeasurable.

1.2 Motivation

Research and development in this area is important because of the nature of the projects [7] volunteer computing supports. These are projects that otherwise may not be commercially feasible right now and such an outcome would hold back the progress of humanity as a whole.

As of this writing, projects powered by BOINC have a total of 2.2 million volunteers between them. A significant portion of the 2.2 million do not contribute with regularity. There are various factors behind this that need to be negated.

1.3 Contributions

Recognizing the potential for bringing more people into volunteer computing and enhancing the experience to make it more meaningful through social networking services, I studied existing work and researched what was possible in this direction.

In time, I architected a system to combine social networking and volunteer computing. This is decribed in chapter 6. The chapter includes details of how the various components including Facebook, the BOINC project servers, BOINC clients and the core Social Cloud application interact.

I introduced a number of social engineering concepts and algorithms to meet the goals of the Social Cloud. They include the concepts of:

- Interest Signature – that helps quantify the interest areas of a user,
- Project Signature – that helps quantify the properties of a project so that users can discover projects that are most suited to their interests,
- Project Champion – that ensures smaller projects get more computational time and support,
- Social Anchor – that helps grow the number of volunteers,
- Compute Magnate – that helps ensure users are socially influenced to contribute with regularity,
- Social value – that helps identify which users are likely to help grow the Social Cloud faster, and,
- Compute value – that helps identify which users might not be achieving their potential to contribute.

They are each described in detail in chapter 5 along with their correlation to the Social Cloud goals. They are later verified to work as intended in chapter 7.

1.4 Publication

Chapters 5 and 6 were largely derived from the paper 'A Social Cloud for Public eResearch' [8]) accepted to the 7th IEEE International Conference on e-Science in Stockholm, Sweden in December 2011. I am the first author of this paper and Kris Bubendorfer, my supervisor, is the second author. Kyle Chard from the University of Chicago is the third author.

Chapter 2

Background

2.1 eResearch

eScience [9] is computationally intensive science that relies on distributed networks to mange large workloads, eResearch extends eScience to other disciplines outside of science. One definition of eResearch describes it as the use of information and communication technologies to enhance new and existing forms of research across disciplines.

eResearch enables research to be done on a scale that was not previously possible, tractable or economically feasible. It facilitates research by making it faster, cheaper and more reusable. eResearch is characteristically collaborative and uses distributed computing technologies to meet the computing requirements that it often requires. For this thesis the term eResearch will be predominantly used but the concepts discussed apply equally to eScience.

A highly regarded example of eResearch would be the Worldwide LHC Computing Grid [10] project. It provides the infrastructure required to process the data generated by the Large Hadron Collider at CERN. To provide some insight into the scale, the data generated annually is over 15 petabytes and is accessible to more than 8000 physicists worldwide. The infrastructure powering this was built by networking thousands of computers housed in hundreds of data centres around the world.

As is expected, given the infrastructure requirements, eResearch is rarely cheap. For scientists, obtaining the funding to get access to such infrastructure is often difficult due to the current structure of funding agencies. There are also licensing and ownership requirements imposed by commercial organizations as a condition to and in exchange for their support.

2.2 High Performance Computing

As the scale and requirements of eResearch far exceed what can be achieved by one or a few personal computers, high performance computing architectures are very important for the purposes of eResearch scientists. In this section, I look at a number of high performance computing models that are capable of meeting the processing and storage requirements of researchers.

2.2.1 Cluster Computing

Cluster computing refers to the use of a group of *tightly coupled* computers that resemble a single computer for large computational workloads. The nodes that form a cluster are often interconnected through high speed local area networks. Clusters deliver higher performance and availability at a lower cost than single computers offering similar capabilities. Many of the supercomputers that rank in the top 500 fastest computers in the world [11] are clusters.

By design, clusters perform better than single computers at handling highly parallel workloads but do not fare as well when it comes to non-parallel workloads.

Of the three different categories of clusters [12] viz. high-availability clusters, load-balancing clusters and compute clusters, high-performance (compute) clusters are the ones most suited for eResearch. While cheaper than a single computer of similar capabilities, clusters also have prohibitive costs associated with building and maintaining them – often running in the tens of millions of dollars for supercomputer scale performance [13].

2.2.2 Grid Computing

Grid computing is similar to cluster computing in that distributed computational resources are combined to process information. The difference lies in the fact that the nodes in grid systems are *loosely coupled* and lend themselves better to processing non-interactive and highly parallel workloads. The nodes are quite likely to be geographically distributed and dissimilar in capabilities.

For eResearch to be feasible on Grid computing systems, given the not inconsiderable cost of Grid infrastructure, researchers would require access to an existing Grid. But access to national Grid infrastructures like TeraGrid [14] and the Open Science Grid [15] is only available to selected projects. There are also strict time and resource limitations imposed due to the limited nature of the resources available in relation to the demand.

2.2.3 Utility Computing

Utility computing involves packaging computational resources and making them available to clients as a metered service. This sets the initial cost of acquiring the infrastructure required by researchers to zero and also lets them scale what is available to them according to their requirements. Utility computing is well suited to researchers who need a high volume of resources for a short period of time. But for long term projects, the costs still remain on the higher side [16].

The underlying computational architecture of utility computing varies with the commercial entity that provides the service. Market leaders in this area include IBM, Amazon, Microsoft and HP.

2.2.4 Cloud Computing

Cloud computing refers to the provisioning of computational resources to clients as a service without them necessarily knowing the physical location of the base resources that fulfil their requirements (hence "in the cloud"). Resources shared can include computational power, storage, software and information services.

Cloud computing [17] encompasses utility computing and grid computing, and shares several characteristics with them. Cloud computing is:

- Agile – It allows for dynamic re-provisioning of resources, thereby improving efficiency of utilization and minimization of costs.
- Reliable – Reliability is a function of redundancy and can easily be factored into a cloud architecture.
- Scalable – The infrastructure is in the "cloud", and can re-provisioned easily and scaled to changing requirements.
- High-performing – Due to the ability to couple computational resources over a network transparently, often using custom middleware and APIs, the capabilities of some of the largest commercial clouds easily meet the reasonable requirements of entities that can afford them.

Cloud computing is very well suited to eResearch. But like utility computing the costs associated with it keep it out of the reach of many researchers [18].

2.2.5 Volunteer Computing

Volunteer computing [19] is a form of distributed computing where individuals share computational resources on their systems with one or more groups of researchers. Volunteer computing came about with the launch of the Great Internet Mersenne Prime Search (GIMPS) [20] in 1996. It gained more traction with the launch of the SETI@Home [21] and Folding@home [22] projects in 1999.

Volunteer computing is highly suitable for eResearch that works with highly parallel CPU-intensive workloads on a tight budget. The processing power made available through volunteer computing matches some of the fastest supercomputers in the world today [11]. Volunteers are usually members of the general public who own computers that are connected to the internet. There are also instances of organizations volunteering unused computational resources.

Volunteers are anonymous and not accountable to projects in this model. However they have to trust the projects that they choose to support - they have to trust that the project is not malicious and won't harm their system(s), and, they have to trust that the project actually does what it claims to do.

Volunteer computing can be considered the most important form of high performance computing because of the cheap processing power that it makes available to eResearch that would not have been feasible otherwise. A study by Berkeley [13] arrived at the following costs for access to 100 Teraflops for a year:

- Cluster computing – $12.4 million
- Commercial cloud computing – $175 million (Amazon's EC2)
- Volunteer computing – $125,000 (BOINC)

All things considered, eResearch projects must be compelling enough to get the general public to support their respective causes and research directions.

Well-known examples of volunteer computing include SETI@Home [21] and Folding@Home [22]. SETI@Home analyzes radio signals coming from outer space in the hope of detecting signatures indicative of intelligent life.

Folding@Home performs simulations of protein folding to provide a better understanding of the development of many diseases. Both projects far exceeded researchers' expectations, gathering huge resource pools and generating worldwide publicity.

2.2.5.1 Volunteer Computing Middleware

Initially, volunteer computing projects relied on a single application for the computation as well as the supporting infrastructure. It was an inflexible approach that resulted in wasted effort and problems with application upgrades.

Recognition of this resulted in the creation of middleware (consisting of client/server software, management tools, web interfaces, etc.) that separated the scientific computation from the infrastructure that supported it. There are a few popular middleware systems in use now including:

- Berkeley Open Infrastructure for Network Computing (BOINC) [23].
- Distributed.net
- XGrid [24] from Apple.
- Grid MP [25] from Univa.

XGrid and Grid MP are both commercial solutions with significant associated costs making them less suitable for cheap eResearch. BOINC and Distributed.net on the other hand are free, but while BOINC is a completely open platform allowing anyone to create projects without passing through any vetting process, Distributed.net only supports a small list of select projects.

The BOINC platform was created as a generic volunteer computing middleware due to the widespread success of volunteer computing projects. As of June 2011, there are over 50 projects powered by BOINC [7], including many that contribute significantly to the global good.

2.3 BOINC

BOINC, the Berkeley Open Infrastructure for Network Computing) is a open software platform for volunteer computing and desktop Grid computing developed by David P. Anderson from the Space Sciences Laboratory at the University of California at Berkeley. BOINC's stated goals [23] were to reduce the barriers of entry into public-resource computing, support diverse applications, to share resources amongst those applications and to reward participants for their contributions.

BOINC had more than double the processing power (5.6 petaflops [4], through its army of volunteers) of the fastest super-computer in the world (Tianhe-I of China with 2.6 petaflops [11]) in March 2011, although there are some inherent performance limitations in the volunteer model [26]. In July 2011 the top ranked super computer reached 8 petaflops, 3 times the power of the now second placed Tianhe-I.

2.3.1 Architecture

In the simplest form, the BOINC architecture consists of clients and project servers with the clients processing data for the projects that they are attached to. The various elements that constitute this basic form are this are shown in Figure 2.1.

2.3.2 Projects

BOINC is best suited for projects that have low data to compute ratios and are likely to have considerable public appeal. There are over 50 listed BOINC projects [7] at the BOINC website covering a wide range of disciplines and research areas. But since BOINC can be used by anyone to create projects, public or private, it is not possible to conclusively determine the total number of active BOINC projects.

An example of a BOINC powered project is Malariacontrol.net, that

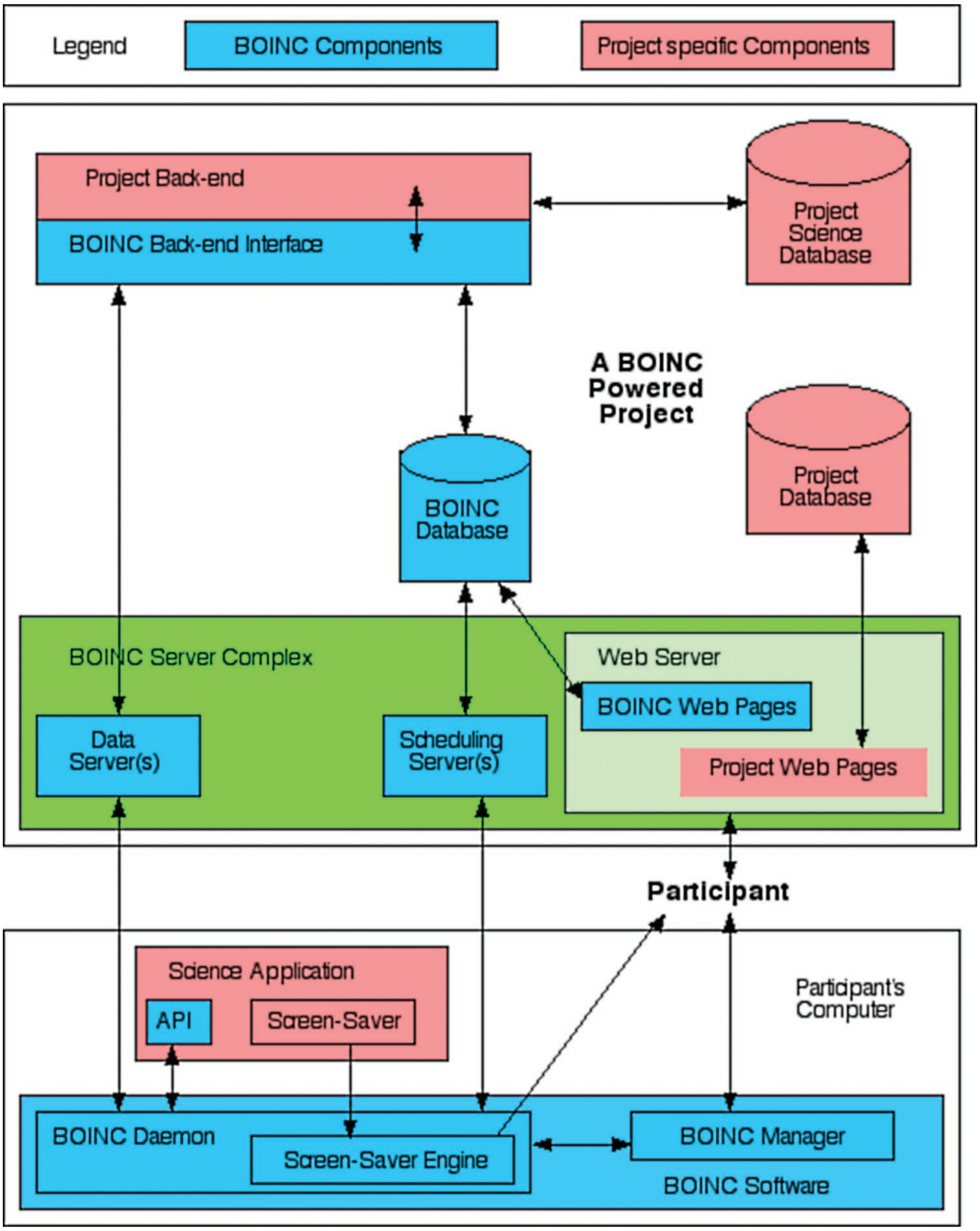

Figure 2.1: BOINC System Architecture [27]

simulates the spread of malaria to determine minimal efficacy and duration of effects needed for a trial vaccine and also to optimize deployment of established treatments. Another is Rosetta@Home, that determines the shapes of new designs for three-dimensional proteins – in order to help find cures for intractable diseases such as cancer and AIDS.

2.3.2.1 Choosing Projects

Given the large number of projects, new users are faced with having to spend considerable time studying what is available to decide which projects to contribute computational time to. This is likely to frustrate and put off many users who are quite likely to be interested in volunteering. It does not help that the project names by themselves are often not descriptive of what they do.

Volunteers are also faced with deciding whether a project is trustworthy – it is entirely possible for a malicious application to masquerade as a genuine project and compromise the volunteer's computer. Even if a project is not malicious, volunteers have to trust that the project follows industry standard security practices to secure their servers.

2.3.2.2 Joining Projects

To join a project, the user has to first create an account with the project. Once the account is verified, the user can *attach* the project to the BOINC client software installed on their machine by providing the project URL when prompted by the client software (section 2.3.3) and then providing their credentials. Minimal user interaction is required beyond this for joining a project.

2.3.2.3 Project Popularity

Project popularity has a direct correlation to the amount of computational power a project sees donated to it. This often unfairly causes biases where older and more established projects see a large portion of the total volunteered computational time go to them, while newer and unknown projects struggle to get the visibility and computational time that they may rightfully deserve.

As the number of projects grows, it becomes harder and harder for the newer projects to get noticed. The unintended consequence of lesser known projects not getting enough visibility is that it might discourage researchers who might be doing potentially groundbreaking research from even trying to utilize the BOINC platform.

2.3.3 Client Software

For a potential volunteer to be able to donate computational time, they must have administrative access to the computer whose computational time they wish to donate, and the computer must have access to the internet. If those conditions are satisfied, they can install the BOINC Client software on their system. BOINC Clients are available for Windows, OS X, Linux and Solaris.

The BOINC client allows the user to 'attach' their computer to one or more BOINC powered projects that they have accounts for. The BOINC client software then downloads work units from the respective project servers and processes them when it detects that the machine is idle. Optionally, a BOINC screensaver can run at the same time. Once the work units have been processed, the BOINC client will send the results to the project servers and claim credit for them.

Users that support multiple projects can adjust the proportion of the total idle computational time that each of those projects receive.

2.3.3.1 Resource Shares

Users can choose to support more than one project by allocating *resource shares* [28] to each project. Resource shares are not percentages, they are a reflection of the portion of the total resource available to a specific project amongst all the projects attached to a given computer. Users use their discretion in determining resource shares and selecting projects. This process may overwhelm less technical users who don't know others who are already contributing to BOINC and are willing to offer advice and help.

2.3.4 Credit System

The BOINC credit system is a means to reward volunteers for the computational time that they have contributed to various BOINC powered projects. The credit system allows users to keep track of their contributions and also compare their contributions with that of other volunteers. It also allows users to compete against each other either individually or in teams. This sense of competition is encouraged to increase the compuational power that is made available to researchers.

When the BOINC client runs on a system, it downloads work units periodically from the various selected projects. It processes these work units, sends the results back to the project servers and claims credit for the work done. Each project has its own specific method of verifying the work done (e.g quorum-based replication) - if the work unit returned is validated, the user receives credit. Credit is not granted if the result is returned after a set deadline or if the result was found to be inaccurate. The credit system is designed to discourage cheating and to encourage users to donate more by creating a sense of competition around credits earned. There are a number of credit statistics sites like BOINC Stats [29] that maintain user rankings.

The validation process prevents credit fraud where users may return bogus data and claim credit. This is unfortunately a reality that needs to be accounted for as there are users who are more interested in public recognition than actually contributing to research.

2.3.5 Account Management Systems

The easiest way for users to manage multiple projects is to rely on an account management system [30]. Account management systems ease the process of joining and contributing resources by allowing the user to set up a 'meta-account' over multiple selected projects.

The account management system creates accounts with the selected projects on behalf of the user. The user can direct the BOINC client on their system to connect to the account management system (rather than a BOINC project) along with their credentials, the account manager then acts as a proxy between the user and the project. They can use the account management system interface to add/remove projects and set resource shares.

BOINC has published a set of WebRPCs [31] that specify how account management systems and project servers should communicate. There are also a set of account manager RPCs [30] specifying how the BOINC clients and account management systems communicate.

Account management systems do not have any social features that would bring in new users and keep existing users involved.

2.4 Social Networks

Social networks (or social networking services to be precise) are online platforms that digitally represent relationships between people, entities that they interact with and activities that they participate in. They then offer services on top of this base data set. Depending on the social network in question, these services can be as simple as basic communications to as advanced and complex as advanced access interfaces for external applications.

Social networking services have seen phenomenal growth in the past few years with it accounting for 1 in every 6 minutes spent online today [32].

There are three major social networking services in the world right now:

1. Facebook – Is the largest social network in the world with over 500 million active users [5] officially. However, the number was recently revealed to be as high as 800 million active users [33]. Facebook supports third party developer applications on its network through a well-documented Application Programming Interface (API).

2. Google+ – Is a fairly new entrant in the social networking space but has seen rapid growth since its launch in June 2011. Google+ did not exist when the Social Cloud for Public eResearch was created. At the time of this writing, Google+ is still in the process of developing and perfecting its API for third party developers.

3. LinkedIn – Is a professional social networking site with over 120 million registered users. It has a documented API that developers can use.

Of all the social networking services in the world, in statistics shared by Neilsen [34], only Facebook ranks in the top 10 web brands in the United States for August 2011 and is only beaten by Google (all websites, not just Google+). It is notable that Facebook led the list in terms of engagement with each person spending nearly 8 hours on average on Facebook in August 2011. The next closest was Yahoo! which saw its visitors spend a little over 2 hours on average.

Social networking is a very broad topic and there is a lot of established work and on-going research in this area. For the purposes of brevity and clarity, only the elements that are relevant to the Social Cloud for Public eResearch will feature in this thesis.

2.5 Facebook

Facebook is a social networking service that was founded by Mark Zuckerberg in 2004 with his roommates and a handful of other computer science students while still at Havard. It was initially open only to Havard students but later expanded to other universities in the United States, then high schools and finally to the rest of the world.

Users on Facebook have a personal profile, can add other people as friends and can communicate with each other publicly or privately. Facebook lets users upload photos, 'like' entities, create/attend events, create/join groups, etc. Additional functionality is made available through applications created by third party developers with specific features for the Facebook platform.

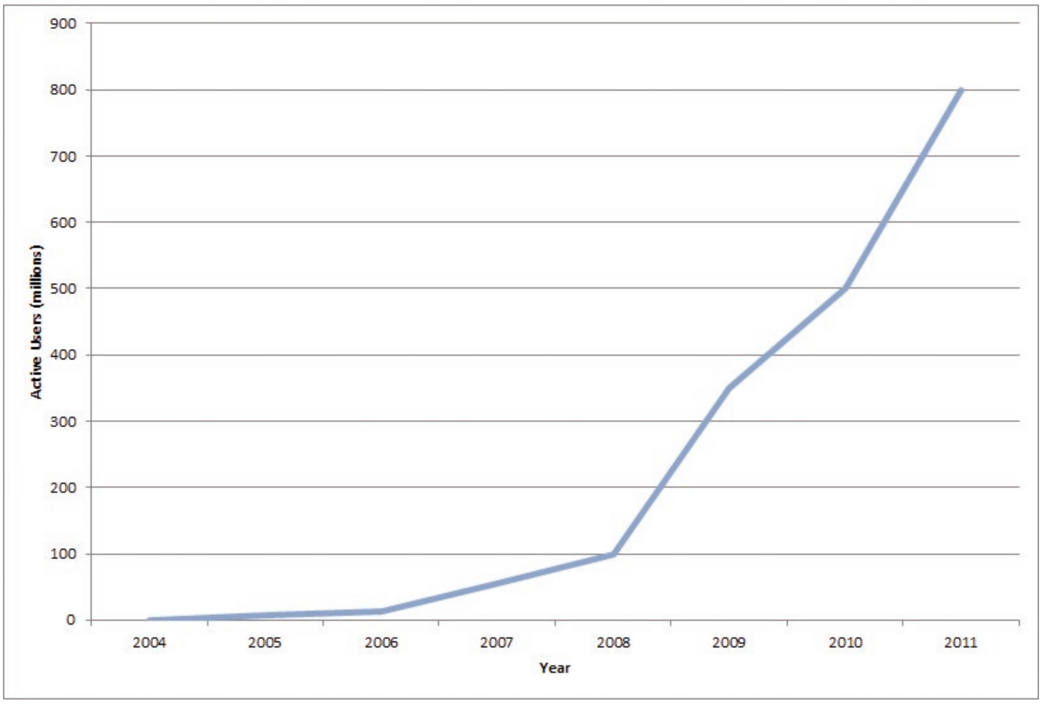

Figure 2.2: The growth of the Facebook user base over the years

Facebook is without a doubt the largest social networking service in the world as can be seen from several independently published metrics in the recent past [33, 34, 35].

Facebook had about a 100 million users in 2008. In three short years, it grew by 700 million users. It's meteoric growth is captured in Figure 2.2. In June 2011, it became the most visited website in the world surpassing 1 trillion page views per month [36]. Statistics published by Citi Investment [35] show the remarkable lead Facebook has over other sites on the internet.

Any effort to utilize the internet as an outreach platform would be ineffective and incomplete without a strong focus on Facebook.

2.5.1 Facebook Development Platform

2.5.1.1 Authentication

Facebook uses the OAuth 2.0 protocol for user authentication, app authorization and app authentication [37]. Each of these processes are supported both at the server side and the client side for third-party applications.

User authentication makes sure that the user is who they claim to be. There are several benefits of this including the possibility of single signons for every third party website and application that decided to support it. A user that logs into Facebook will automatically be identified to any application that they've authorized when they access its URL. The convenience that this brings about reduces the effort required of a user to register with and log in to a third party application.

App authorization ensures that the user is in control of their information with respect to third party applications. The user is explicitly told what data that the application is asserting that it requires to work properly, and the user can either decide to allow the application access or to cancel any further interaction with the application. Possible user data permissions that an application can request from the user through Facebook include *user_about_me*, *user_birthday*, *user_education_history*, *user_location*, *email* and many more [38].

App authentication ensures that user data is only sent to the application that the user has authorized and not to any other.

2.5.1.2 Graph API

The Graph API [39] is what allows third party developers to read and write data to Facebook. It enables them to interact with the objects and connections that form the *social graph*. Objects in the social graph include people, photos, pages and more.

Privacy and security being important, access to objects in the graph is restricted to applications authorized to access those objects. The first name, last name and profile picture of a user are freely accessible for example, but access to more personal details like the user's date of birth or private photos would require permission from the user concerned.

Actions possible with the Graph API include reading, publishing and deleting of objects.

2.5.1.3 Open Graph Protocol

Facebook's Open Graph protocol [40] is designed to help represent real world entities as objects on Facebook. This enables Facebook users to interact with those entities, and Facebook can realize and represent these interactions as an integral part of the Facebook experience. Every object in the social graph has a unique identifier associated with it and this unique identifier is used for any sort of manipulation of data associated with that object.

For third-party developers, using the Open Graph protocol is vital to integration with the main distribution points in Facebook - namely the News Feed, Requests, Notifications and User Profiles. In late 2011, additional points of integration opened up with a new version [41] of the Open

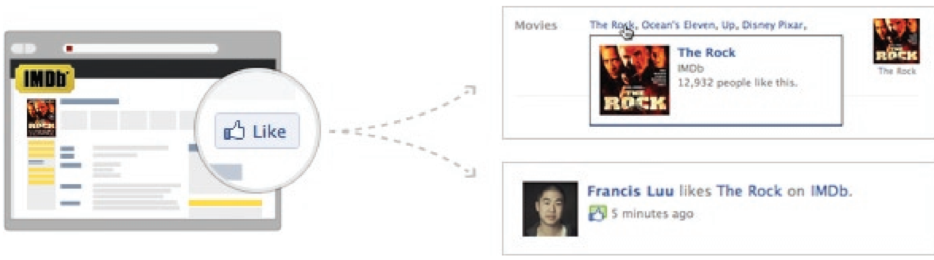

Figure 2.3: Utilizing the Open Graph protocol [40]

Graph protocol but that is out of the scope of this thesis.

2.5.1.4 Social Channels

One of Facebook's biggest draws for third-party developers is the ability to leverage the social graph to potentially virally transmit information through users and their connections.

The News Feed is a personalized ever-updating view of what a user's friends are doing on Facebook or in authorized applications. Data published on to this feed is aggregated from data published to the *walls* of each of the user's friends. Facebook has algorithms in place to decide the relevance of those pieces of data so that a user's news feed contains information that the user is likely to find most interesting.

Requests are notifications that users receive from their friends asking them to take specific actions in a third-party application. Requests can either be generated explicitly by the user or automatically by an application that has been authorized by the user to do so.

There are also a number of *automatic* channels like bookmarks, notifications, dashboards, usage stories, and app profiles & search [42].

Chapter 3

Related Work

3.1 BAM! (BOINC Account Manager)

BAM! [43] is a website that lists available BOINC projects and aids users in finding ones that they maybe in interested in contributing to. It works as a full account manager with data aggregation and management from multiple BOINC project websites. It allows users to centrally manage all their project accounts and remotely manage the BOINC client on their machines. BAM! allows user to create or join teams that compete against each other on the basis of credits earned.

It is important to note that BOINC Account Manager (BAM!) is a just one of a handful of popular BOINC powered account managers (section 2.3.5) and it is not to be confused with the entire class of such applications. BAM! improves upon the standard BOINC account manager through BOINC Stats.

3.2 BOINC Stats

BOINC Stats [44] is the larger website that hosts BAM!'s web interface. BOINC Stats performs serveral functions:

- It hosts a web forum and fosters a sense of community among BOINC participants.
- It gathers project data from individual project servers and compiles statistics for public perusal.
- It lists project challenges and team challenges that users can work towards.

3.3 GridRepublic

GridRepublic [45] is another popular BOINC account manager that was developed in close coordination with BOINC. GridRepublic is a community effort with donations and volunteers contributing to the upkeep of the website.

Like BAM!, GridRepublic supports a large number of BOINC projects [46] and helps users find ones they are interested in by maintaining a consolidated list of projects. It also has a community forum [47] where members can discuss volunteer computing related efforts and best practices in growing the user base of GridRepublic and BOINC.

3.4 Progress Thru Processors

In 2009, BOINC has collaborated with Intel and GridRepublic to create a Facebook application called Progress Thru Processors [48].

The Progress Thru Processors Facebook application provides a streamlined way to create a GridRepublic account from within the Facebook platform. While there is some account management functionality within the Facebook application, users are redirected to GridRepublic to perform most non-trivial tasks. Overall, there is still a lot of scope to integrate volunteer computing models into Facebook, and leverage social connections between users and their associated incentives.

Progress Thru Processors has seen modest success, however we believe there is much greater success to be attained if the social aspects of social networking are more tightly integrated.

3.5 Social Cloud

A Social Cloud is a scalable computing model in which heterogeneous resources contributed by users are dynamically shared amongst a group of "friends" in a social network. A Social Cloud benefits from an implicit level of trust between users and the associated socially corrective mechanisms that exist due to the real-world basis of the relationships represented. The cloud-based usage model enables virtualized (elastic) resource sharing through service-based interfaces exposed by members of the network.

One way of thinking about the Social Cloud is to consider that social network groups are analogous to dynamic Virtual Organizations (VOs) [49]. Groups, like VOs, have policies that define the intent of the group, the membership of the group and sharing policies for the group.

This model is illustrated in Figure 3.1, where user-specific groups, defined by relationship types, are shown in the context of a social network. In this example group A is composed of only co-worker members, whereas group B is formed by family members and group C includes only friends. Clearly the level of trust and mechanisms for *social correction* (identifying incentives and disincentives for users to participate) differ between groups. This figure also highlights that social clouds are not mutually exclusive, that is, users may be simultaneously members of multiple social clouds. Whereas a VO is often associated with a particular application or activity, and is often disbanded once this activity completes, a group is longer lasting and may be used in the context of multiple applications or activities.

More recently *crowdsourcing* has emerged as a hugely successful dis-

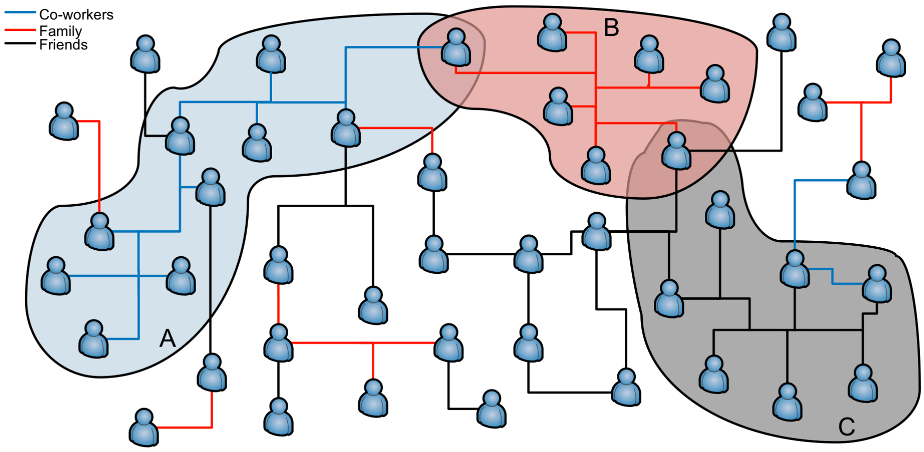

Figure 3.1: Social Cloud overlay in a Social Network [6]. Three different Social Clouds are illustrated to highlight the use of relationships when establishing Social Clouds.

tributed problem solving model in which large scale tasks are broadcast to, and solved by, an unknown group of amateur members of the public [50]. This is a more general model of volunteer computing in which the donated 'resources' are the contributors' personal skills and their time.

Chapter 4

Application Design

The goal of the research is to study ways in which social networks, specifically Facebook, can be integrated with volunteer computing middleware, specifically BOINC, to effectively raise the computational resources available to researchers.

4.1 Reason/Need

Despite the notable success of BOINC, it has primarily relied on world-of-mouth publicity. There are also perceived (and real) barriers to entry for less-technical users. For example, it is difficult for a non-technical user to understand BOINC, discover projects, and install a BOINC environment for contribution [51]. Nearly all current barriers stem from the fact that BOINC was originally created by, and for, a technical and knowledgeable audience.

BOINC took a major step towards addressing these barriers by collaborating with account management systems [30] like GridRepublic [45] to ease the management of multiple projects for volunteers. Before the introduction of account management systems, users who wanted to support multiple projects had to manually setup accounts with each of the projects and then manage contributions separately. This was a source of frustration for early users.

4.2 Case Study

An excellent example that shows the benefits of introducing social networking elements into a system can be seen in the comparison of Electronic Arts and Zynga.

Electronic Arts (EA) is a major North American video game company that was founded in 1982. It develops, markets and distributes video games around the world. It had revenues of around $3.6 billion and was worth around $4.6 billion [52] in 2010.

Zynga is a social network game developer that was founded in just 2007. As of 2011, Zynga's Facebook games see over 230 million active users [53] every month. In 2011, Zynga was valued as high as $20 billion dollars [54].

Electronic Arts has, thought late, realized the opportunity that it was missing and shored up its social networking offerings and has since become the second largest social game developer on Facebook.

4.3 Stakeholder Motivations

The Social Cloud for Public eResearch represents a unique computing environment in which users of a social network are able to donate resources to scientific computing projects. In doing so, this work aims to take volunteer computing from more technically oriented users towards everyday users of a social networking service.

The adoption of Social Cloud computing provides a novel mechanism for leveraging social engineering to motivate and facilitate volunteer based sharing. Specific motivations for users and researchers are detailed in the following sections and referenced throughout the remainder of this paper.

4.3.1 Users

There are several reasons users participate in volunteer computing projects, in general users are motivated by altruistic or self-interested reasons. For example, altruistic users may have a desire to make a difference or have a strong interest in a particular field of research, whereas self-interested users are typically motivated by competition with regards to the contribution size (leaderboards) or a desire to be publicly recognized.

It is reasonable to assume that only a very small percentage of the general public is interested in volunteer computing. Generally, the projects themselves drive a user's desire to contribute, and in essence volunteer computing is just a means to an end for these users. Specific motivation of volunteer computing users have been studied in [55].

Briefly, the following key factors were identified as strong motivation for volunteers:

- the potential impact of the science,
- the probability of success,
- the utility of the project,
- the safety of the project
- the political signals that the project sends out,
- the democratisation of science, and,
- personal benefits such as a sense of community, competition, personal interest and visual pleasure.

In current volunteer computing initiatives users are responsible for finding appropriate projects, weighing up which projects are most suited to their interests, and setting up and maintaining required volunteer software. Clearly this is a barrier for some, perhaps many, users. In a social cloud relationships between users can be used to share information and determine suitable projects for participation, while resources can be shared more easily through a simple social network application.

The social relationships defined in a Social Cloud present a unique way of providing many of these key motivating factors. Specifically targeted social algorithms and techniques can be used to both maximize the number of new volunteers, and, to keep them engaged and involved so that the computational time available to researchers grows quickly and sustainably. The Social Cloud provides a single integrated management view (within the social network) of all projects a user contributes to to easily monitor current activity and also explore new projects.

4.3.2 Researchers

Project owners face similar challenges as they must advertise their projects, and, determine and implement appropriate motivation mechanisms to encourage user contribution. In addition to benefiting users, the Social Cloud for Public eResearch aims to provide increased publicity to different volunteer projects.

Through the Social Cloud researchers will be able to connect with the people (and groups) that support their research. It is therefore beneficial for all parties as research can truly become participatory and inclusive. Significant public interest in a project can result in new support and interest from external sources of funding. It can also be a tool to influence political interests in areas of research that polarize public opinion. It is conceivable that an increase in the number of volunteers may unintentionally favour high visibility projects and therefore discourage researchers from trying to exploit volunteer models for newer less established projects. It is therefore very important to ensure that such projects are not disadvantaged through appropriate algorithms and policies.

In addition to the benefits described, the Social Cloud model lowers the barriers of entry both for volunteers and researchers through its resource sharing framework, thereby providing a large amount of processing power that would otherwise be irrecoverably wasted.

4.4 Requirements

From the case study and studying the various stakeholder motivations in detail, the following requirements were derived:

1. Ease the process of an interested party becoming a volunteer.
2. Enhance the visibility of lesser known BOINC projects.
3. Incentivize user involvement, contributions and platform growth.
4. Maximize the computational power available to researchers.
5. Bring science closer to the general public and make it more meaningful.

In the subsequent chapters, the various ways in which the Social Cloud for Public eResearch meets these requirements will be elaborated, substantiated and validated.

Chapter 5

Social Engineering

Unlocking the power of social networks requires delving into social engineering, for example motivating user behaviour based on social incentives. Current volunteer platforms like BOINC do not explicitly consider social engineering. However, it is our view that social engineering should be considered an important factor in volunteer computing. The underlying social network in the Social Cloud provides deep access to the social relationships between users and therefore an opportunity to exploit social mechanisms to encourage and maintain contribution.

A key aspect of social engineering with respect to the Social Cloud is using facets of viral marketing in reaching out to new users. Viral marketing is a marketing technique that relies on a self-replicating process to sustain it – not much unlike a virus. Viral marketing has been found particularly effective on the internet and lessons from this are in play in the Social Cloud for Public eResearch.

5.1 Viral Marketing

Marketing is essential to enhancing the visibility of BOINC projects. However, financial constraints faced by researchers greatly restrict the options available to them when it comes to marketing their projects. Viral marketing is perfectly suited for situations like this since if it is properly implemented, it is self-propagating and grows automatically as a matter of course.

I considered the key strengths of successful viral marketing campaigns, both on the internet and in the real world. I independently came up with several concepts that are described in detail in section 5.4. My conviction in those new concepts got reinforced when work in a similar direction was published by marketing professors Andreas Kaplan and Michael Haenlein [56] in May 2011.

According to them, the attribute that is common to all successful viral marketing campaigns is that they all give the "right message to the right messengers in the right environment".

The **message** necessarily has to be memorable and interesting enough to trigger an individual to want to share it with their friends. The Social Cloud is well positioned to control the narrative by suggesting powerful and compelling pitches to users to share with their friends on Facebook. This is an improvement over users having to come up with their own messages that may strictly be a hit or miss affair when it comes to effectiveness. That said, users will be able to tailor the suggestions the Social Cloud gives them to make the message personal.

The **environment** is also controlled as most of the messages will be within the context of a social network and integrate the pull afforded by a personal message from a friend with whom the potential target has an existing relationship. Being inside a social network also makes the message more likely to go viral as the logical structure mirroring real world relationships for message transfer is already in place.

The most important factor that determines the outcome of a viral marketing campaign is the **messenger**. Andreas Kaplan and Michael Haenlein identify three types of messengers that play important roles in turning an ordinary message into one that goes viral. They are market mavens, social hubs and salespeople.

Market mavens are portrayed as individuals who are on the cutting edge, who are traditionally *first adopters* of new technologies and concepts. Market mavens represent the starting point of a viral marketing campaign. In the context of the Social Cloud for Public eResearch, they are quite easily identified. They are existing users of the BOINC client who are passionate about contributing computational power. They are usually the ones that discovered BOINC with minimal exposure to a coordinated viral marketing campaign.

Social hubs are people with an unusually large number of friends. They are outliers when it comes to the number of people they are connected to. They are well positioned to act as a means for market mavens to send out their message to a vast number of people. They are easy to identify in the context of the Social Cloud because of the data available to the us from Facebook.

Salespeople are those that serve to make the message relevant and compelling before passing it on to others. They usually find their place between the market mavens and the social hubs. While they form a separate category in traditional viral marketing, the Social Cloud attempts to impart every user with attributes of salespeople by controlling and helping shape the narrative, as mentioned earlier. In a sense, the Social Cloud is a highly influential pseudo-salesperson that strengthens the link between every market maven and social hub that are connected to each other.

That concepts similar to those employed in the Social Cloud for Public eResearch were created independently by other researchers at almost the same time is indicative of their soundness. They are further described in section 5.4 below.

5.2 Gamification

After getting people interested in contributing computational time to BOINC projects, one of the most important goals of the Social Cloud for Public eResearch is to keep them contributing and helping meet other goals. Passively relying on their self-motivation to assist us in meeting those goals is unrealistic.

We attempt to motivate individuals by tapping into their sense of competition, by playing to their desire for public recognition and by relying on social pressure from their friends. This is also known as *gamification* - the process of using mechanisms typically found in games to induce desired behaviours in a non-game application.

Like viral marketing, gamification is a process that has been proven to work time and again if implemented correctly. There are even companies like Badgeville [57] whose business model revolves around helping non-game application companies gamify their offerings. That gamification is effective is evidenced by Gartner's prediction that by 2015 over 50 percent of companies will embrace gamification [58].

Appealing to a user's sense of competition and their desire for public recognition has traditionally involved individual leaderboards on BOINC statistics sites [29]. However, this is restricted to the credits earned and there is no real incentive to bring in new users. The exception is when teams are established [59] as the credits earned are pooled. This has proven to be a successful strategy for growing computational contributions, however, additional potential lies in tying it to social networks and extending it to more than just computational credits.

Given the social context created by the Social Cloud, users are able to compete directly with their friends - people they have far stronger social relationships with. This mechanism acts both to encourage increased individual contribution but also to encourage new friends to join in.

5.3 Motivation Theory

In order to better align and link the motivations of the Social Cloud with those of the targeted end-user, we attempted to compare the means we intended to employ with past work on human motivation theory, notably work by David C. McClelland in Human Motivation [60].

It is asserted that human motivation is dominated by three basic needs – the need for achievement, the need for power and the need for affiliation.

The *need for achievement* is suitably met through the use of gamification. This need is met by the Social Cloud by presenting goals tailored to the user that are challenging yet realistic. This is a marked improvement over the previously mentioned BOINC statistics sites [29] that favour long term contributors who joined early rather than new contributors who would have to struggle to catch up.

The *need for power* is the desire to lead and make an impact. This need is well suited to getting people to bring their friends on board as contributors to help make a difference with the research they support.

The *need for affiliation* can be better described as a need to feel liked and accepted. This is a powerful need that can be harnessed very well in the setting of a social network like Facebook where actions by users can be shared and commented upon by friend connections.

The specific ways in which these known needs are used are explained in section 5.4.

5.4 Social Cloud Concepts

Building on the concepts touched upon in sections 5.1, 5.2 and 5.3, I created a number of structured incentives to work towards the various goals of the Social Cloud for Public eResearch. The incentive mechanisms themselves and how they come together to enable us to achieve various goals are detailed in section 5.5.

At this point, it is important to distinguish between *the Social Cloud*, which is the overarching reference to the application itself and the various actors that actively participate, and *a user's social cloud*, which is the set of their friends that are members of the Social Cloud.

5.5 Social Cloud Concepts in Action

5.5.1 Easing the Process of Joining

When a potential user arrives at the Social Cloud Facebook application, there are a number of steps that they will go through before they become valuable volunteers to BOINC:

1. Understanding how volunteer computing works and why it is important.
2. Selecting projects to support.
3. Choosing resource shares.
4. Installing the BOINC client and configuring it properly.

The first step is critical for capturing the user's interest and keeping them motivated through the subsequent stages. From a social engineering perspective, this can mean including compelling hooks from their friends and providing various *calls to action*.

5.5.1.1 Interest Signature

Given that BOINC projects are quite numerous and cover a number of research areas [7], I came to the conclusion that there needed to be a way to help users pick out projects that they might be interested in. But for this to be possible, it was necessary to create a means to reliably quantify a user's areas of interest and the concept of an *interest signature* came about.

An interest signature defines a point in n dimensional space describing an individual user's specific areas of interest with each dimension quantifying a well-defined field of interest. Each dimension represents one of several areas of interest including, but not limited to, mathematics, cryptography, climate study, biology, physics, astronomy, chemistry and artificial intelligence. A generic representation of the interest signature, I, of a user, u, in n dimensions would be as follows:

$$I_u[n] = (i_0, i_1, i_2, i_3, ..., i_{n-1}) \tag{5.1}$$

The interest signature is obtained from the user by explicitly asking them to rate their interest in each predetermined research area on a scale of 0 to 10 with 0 representing no interest and 10 representing maximum interest. A user with a strong interest in mathematics might indicate 10 in mathematics, 8 in physics, 8 in cryptography, 2 in biology and so on. A user interested in helping cure diseases and in ensuring a better world to live in might rate climate study and biology highly.

The interest signatures that are obtained from users are normalized to enable reliable comparison. The Social Cloud for Public eResearch calculates the interest signature distance, D_{uf}, as defined in equation (5.2), between a user, u, and all their friends, U_f, based on their interest signatures, I_u and I_f, to identify friends with the most similar interests to that particular user. The shorter the Euclidean distance between two interest signature points, the more similar the interests of the two users.

$$\forall f \in U_f, D_{uf} = \sqrt{\sum_{i=0}^{n-1} (I_u[i] - I_f[i])^2} \tag{5.2}$$

5.5.1.2 Project Signature

Given that an interest signature is a representation of the areas of interest of a given user, there had to be an equivalent representation for projects. Projects can be considered to have a mix of different *areas of appeal*. Areas of appeal have a one to one correlation with the set of areas of interest in the Social Cloud and are essentially one and the same except for their application.

Like interest signatures, a project signature defines a point in n dimensional space describing the areas of interest that a project encompasses with each dimension quantifying a well-defined area of interest. A generic representation of the project signature I, of a project, p, in n dimensions would be as follows:

$$I_p[n] = (i_0, i_1, i_2, i_3, ..., i_{n-1}) \tag{5.3}$$

There were two methods considered to populate a project's signature. One would be to depend on subjective assessments of either the administrator(s) of the Social Cloud or the administrator(s) of the individual projects. This would be necessary when the very first member of the Social Cloud is yet to join or when the number of users in the Social Cloud contributing to the given project is low.

The second way to do it would be to base it off the interest signatures of the set of users that contribute to the project in question. A simple average of each interest dimension resulting in the corresponding appeal dimension would be a close approximation. But then that would give equal weightage to the interests of people who might donate 10% and people who might donate say 90% of the total free computational time on their machine. It is obvious that an unweighted average could give a misleading picture of a project's areas of appeal and consequently the project signature itself. So weights proportional to each user's percentage resource share in the project under consideration are applied.

That said, it is always possible that malicious user accounts or misinformed users might skew a project's signature. Given a sufficiently large pool of users, the effects of this are expected to be minimal. But further investigation and research is needed in this direction and that is out of the scope of this thesis.

5.5.1.3 Signature Distance and Project Selection

By calculating *signature distances*, we are able to highlight projects that were chosen by users, either globally or within the user's set of friends, that have interest signatures similar to the new user.

We calculate the signature distance, D_{up}, between a user's interest signature, I_u, and a project signature, I_p, as follows:

$$D_{up} = \sqrt{\sum_{i=0}^{n-1} (I_u[i] - I_p[i])^2} \tag{5.4}$$

This enables us to help the user easily pick projects that appeal to them.

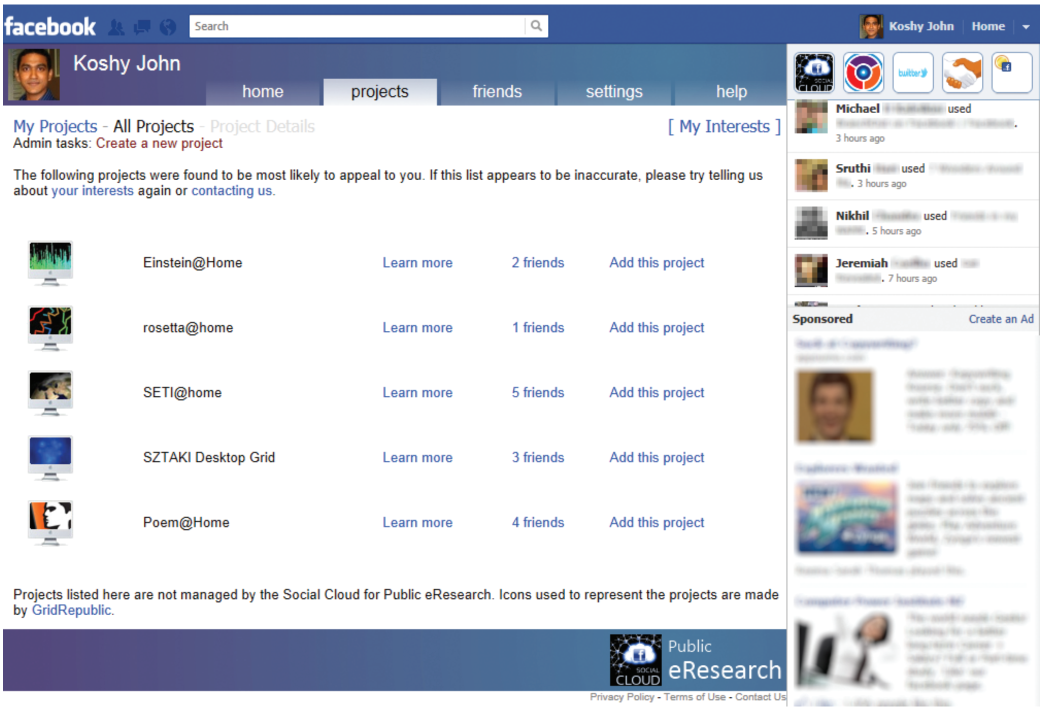

Figure 5.1: A screen shot showing the project suggestions based on interest and project signature distances.

Figure 5.1 shows a screen shot of the prototype, in this figure the user is shown a list of BOINC projects that are most suited to their interests. This list of projects are selected based on the distance between the project signature and the user's own interest signature.

Users are also shown the number of their friends that support each project and may select projects and shares from this view. Additionally, new users that are not sure of which projects to choose can align their choices with those of their friends. Users that want a reliable personal opinion on a project are directed to *project champions* (subsection 5.5.2.1) within their group of friends.

Once users select projects, they are required to select individual resource shares for each of the projects. Recall, a resource share is a reflection of the percentage of the computational time that a project gets on a user's machine. Users can select their own shares but we also provide recommendations based on averages of the normalized resource shares of their friends. This helps both reduce technical barriers to entry by enabling users to make meaningful resource share choices when in doubt. It also helps catalyze engagement by encouraging competition – as users will be evenly matched with their friends with regards to resource share distribution among projects.

Finally, users are given instructions on installing and configuring the BOINC client, and are directed to their friends in the event that they need assistance at any point.

5.5.2 Incentivizing Involvement, Contribution and Growth

We have developed mechanisms akin to gamification to ensure that users contribute to growing the user base and the computational time available to BOINC projects. Scores related to various goals are calculated periodically to establish user rankings and to identify the best contributors within each user's set of friends.

We have introduced three high scoring sets of users - *project champions*, *social anchors* and *compute magnates* based off *project credits*, *social scores* and *compute scores* respectively. These labels are local to the view of each user and are reserved for friends that fall into the top bracket in a particular score. Project credits are routinely queried from project servers while social scores and compute scores are periodically recalculated using algorithms detailed in the following sections.

To motivate users and to enable them to measure their progress we utilize leaderboards specific to each user's personal social cloud. Significant changes in positions on the leaderboard are published to the user's Facebook wall – this in turn is expected to create interest in the application in addition to giving the user a sense of recognition.

There will be situations where a there might not be enough data to work with, like when a user with no friends joins the Social Cloud. In those situations, we fall back to values based off the Social Cloud as a whole.

5.5.2.1 Project Champions

Project champions are considered to be the "biggest" contributors to a given project and therefore represent the best go-to person for a user from within their friends list if they need to know more about a project. Due to their high contributions they are expected to champion the cause of the project. Project champions are identified based on the total credits that they have earned contributing to a given project.

Project credits are routinely queried from project servers corresponding to each project. Since BOINC has mechanisms in place to ensure that people get credit for only valid results submitted by their machine(s) [61], ensuring that project credits were obtained through genuine contribution.

It is easier to become a project champion of a less popular project, and we therefore expect the desire to become a project champion to also increase the computational time that smaller projects receive. This is an important goal of the Social Cloud (section 4.4).

5.5.2.2 Social Anchors

Given the underlying social nature of the Social Cloud, it is expected that the majority of users will be introduced to it by their friends either explicitly (through an invitation or status message on their Facebook wall) or serendipitously where a potential volunteer may chance upon a report of the contributions of their friends in their Facebook news feed.

Existing Social Cloud users are incentivized to bring in their friends into the Social Cloud by tying actions in this direction to a *social score*. The top bracket of friends in terms of social scores are identified as *social anchors* in the user's personal social cloud. Breaking into that top bracket in their social cloud earns the user the title of social anchor.

Social Scores

$$\forall u \in U, Sv_u = \frac{n_f + 1}{n_{scf} + 1} \tag{5.5}$$

$$\forall u \in U, Ss_u = \sum Sv_{uf} \tag{5.6}$$

A user's social score represents a measure of their continuing contributions to the growth of the Social Cloud. Social scores are used to motivate existing users to encourage their friends to join the Social Cloud. New users with the least number of existing friends in the Social Cloud are of high value because they are less likely to have joined otherwise and have high potential to bring in new users, the users that are responsible for bringing them in are appropriately rewarded with a higher boost to their social score. New users with a lot of existing friends in the cloud are of less value because the effort required to get them to join is likely to be lower. The increase in the social score of the users associated with the new user would be consequently lower.

To this end, every user has an associated social value, Sv_u as defined in equation (5.5). The higher the number of friends, n_{scf}, they have *in the Social Cloud* the lower their social value. The higher the number of friends they have *in total in the social network*, the higher their social value. Social values of a user's friends (in the Social Cloud), Sv_{uf}, add up to give the user their social score, Ss_u as defined in equation (5.6). For a user to maintain a high social score, they would need to keep recruiting users who are less likely to join.

The social value of users who have increased the number of friend connections in the cloud since they joined will decay by virtue of that fact. This serves the dual purpose of disincentivizing users adding existing Social Cloud users as friends on Facebook to boost their social score and to ensure that users don't rest on their social score achievements.

Social anchors are the key to growing the pool of users (a requirement of the Social Cloud mentioned in section 4.4), and the title is recognition for their continuing contributions in this direction.

5.5.2.3 Compute Magnates

Compute magnates represent the top bracket of friends generating *valuable* computational time for the Social Cloud through their *own* social clouds. The computational time is based on the calculation of individual compute scores. This title was shaped to serve the dual purpose of incentivizing higher computational time contributions and application of social pressure on friends to maintain or improve on their contributions.

Compute Scores

$$\forall u \in U, Cr_u = \sum C_{u30} \tag{5.7}$$

$$\forall u \in U, Cv_u = \frac{Cr_u}{n_s cf + 1} \tag{5.8}$$

$$\forall u \in U, Cs_u = \sum Cv_{uf} \tag{5.9}$$

Compute scores are a reflection of how much credit a user and their set of friends generate in a rolling 30 day window, C_{u30}. They are used to encourage users to ensure that their friends are contributing computational time with regularity. Like social scores, users are incentivized to focus on people with fewer friends in the Social Cloud, and disincentivized from adding friends simply to boost their compute scores.

The rolling credit value of a user, Cr_u, is used to generate their compute value, Cv_u, by dividing it by the number of friends, $n_s cf$, in their social cloud. Compute scores, Cs, for every user are then generated by summing up the compute values of all their friends as shown in equation (5.9).

There is an alternative to using the rolling credit value I have defined here. It is to use BOINC's own Recent average credit (RAC) [62] for every project that the user contributes to and then average them out again to obtain a single value. Recent Average Credit (for a single project) is obtained by taking the total credit and halving it every week before summing it with the latest granted credit. In effect it is supposed to reflect the rate of your contributions for a day on average in the recent past. However, according to BOINC [61], there are many factors not considered including host processing inconsistency, delay in work unit validation and project down time. It is in this light that we decided to stick to the rolling credit value in determining compute scores.

Because of the method of calculation, users who have friends that contribute less than what they (the users) stand to gain in terms of their own compute value may feel incentivized to remove those friend connections. But this would work only through breaking the Facebook friend relationship itself and *we feel that most relationships are strong enough for the user to work on getting their friend to contribute more instead*.

Chapter 6

Architecture

The Social Cloud for Public eResearch can be visualized as shown in Figure 6.1. It is essentially a privately hosted multi-faceted application designed to work with BOINC and Facebook.

From the perspective of project servers and volunteer PCs running the BOINC client, the Social Cloud for Public eResearch is an account management system. From the perspective of users (volunteers), the Social Cloud is a socially aware account management system which runs as a Facebook application. The existence of the Social Cloud is transparent to researchers and requires no additional effort on their part to support.

As discussed previously, the Social Cloud considers a number of factors that may prevent a new user from engaging in volunteer computing, and attempts to address those by tapping into their social circles.

Users can add and remove supported projects from within the Facebook application, they can also set relative resource shares (percentages of the total computational time donated) for the various projects that they choose to support. This information is used to communicate with the various project servers and to control the BOINC client running on the user's PC.

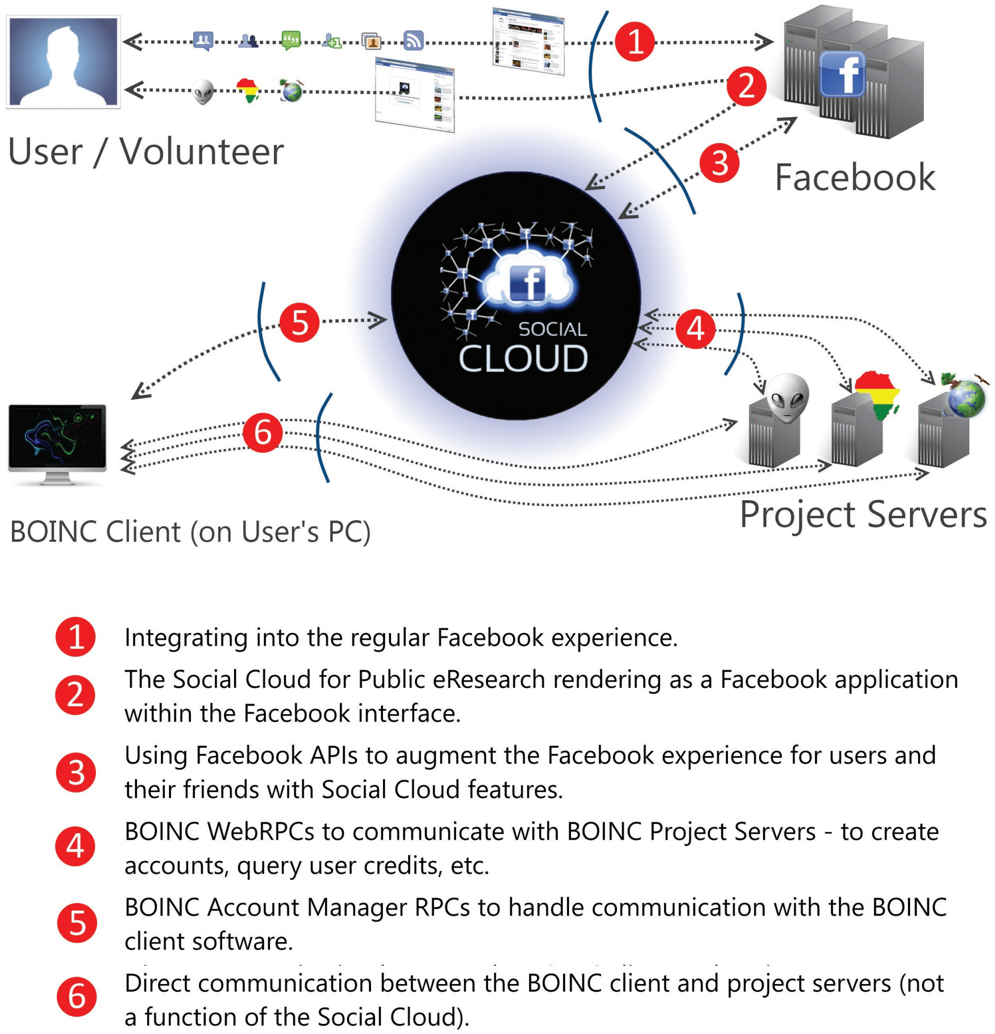

Figure 6.1: Architecture of the Social Cloud for Public eResearch

6.1 Facebook

The Social Cloud is built on Facebook for reasons detailed earlier. Facebook allows externally-hosted applications to run transparently within the Facebook UI. Access to social information is provided through the Facebook Graph API [39].

The Graph API exposes access to the underlying social graph that contains users and their connections with other nodes in the graph (people, photos, events, pages, etc.). To access the Graph API, both the user and the application must be authenticated by Facebook using the OAuth 2.0 protocol [37]. The power and potential of the Graph API combined with the vast user base that Facebook has made it the obvious social network for this project.

The social data required to provide a meaningful experiences for our users is obtained through the Graph API. We also extend the Social Cloud experience back deep into Facebook by manipulating Facebook objects in the same manner.

6.2 BOINC Project Servers

As far as BOINC project servers are concerned, the Social Cloud for Public eResearch is just an account management system. As long as a project supports BOINC's published Web Remote Procedure Calls (WebRPCs) [31], the Social Cloud account manager can support it.

The WebRPC model assumes every RPC to be an HTTP GET transaction, the input parameters are represented as a set of parameterized GET arguments. The resultant output is an XML document with well-defined fields that is parsed by the social cloud to let users monitor their contributions and also to feed our social engineering algorithms described in the previous section.

6.3 BOINC Clients (on Volunteer PCs)

Locally deployed BOINC clients can be attached to an account management system in various ways. Data about the account manager can be bundled with the installer, or the user can specify the account management system URL (in this case, it is the URL to the Social Cloud for Public eResearch). The user will provide authentication details for the BOINC client to obtain their project and resource share preferences from the Social Cloud.

The BOINC client communicates with the Social Cloud account management system using Account Manager RPCs [30] published by BOINC. Once the client has processed data relating to the projects that the user supports, it attaches itself to each of the project servers directly and starts pulling information for processing.

6.4 Interactions

To help understand the architecture we describe a basic usage scenario for the Social Cloud through a sequence diagram in Figure 6.2.

The process starts when a Facebook user discovers the Social Cloud for eResearch application. If they choose to add the application, permissions for the required user data are requested through Facebook.

Once the application has been added, the user is presented a form to determine their interests in various research areas – this is used to generate their interest signature. The interest signature generated is then compared against the project signatures of all the available projects to determine appropriate projects ordered (by projected interest) for this user. The user can then select any projects that appeal to them and the Social Cloud proceeds to create accounts for them at each of the various BOINC project servers on their behalf.

Suggestions are made to the user on the resource shares that they should allocate to the various projects that they have selected. These suggestions are based on the normalized resource share values of their friends. This allows for meaningful competition in the future.

The user is then prompted to install the BOINC client and provide credentials to connect to the Social Cloud. The BOINC client pulls information regarding the projects that the user has selected along with their resource shares from the Social Cloud. It connects to each of the project servers and downloads work units for processing. In due course, the results of the processing are sent back to the project servers. Each project server verifies the results obtained and grants credits as appropriate.

The Social Cloud routinely queries credits for every user from individual project servers. If a user is found to have achieved a credit milestone in a project, it is published to their Facebook wall. This is visible to friends and should generate cascades of interest. The user can also view the application at their convenience to check on their progress and that of their friends. They may also suggest the Social Cloud for Public eResearch to their friends to help drive its growth (and increase their social score).

In addition, the Social Cloud periodically processes data available to it to establish rankings and achievements for users based on the algorithms described previously.

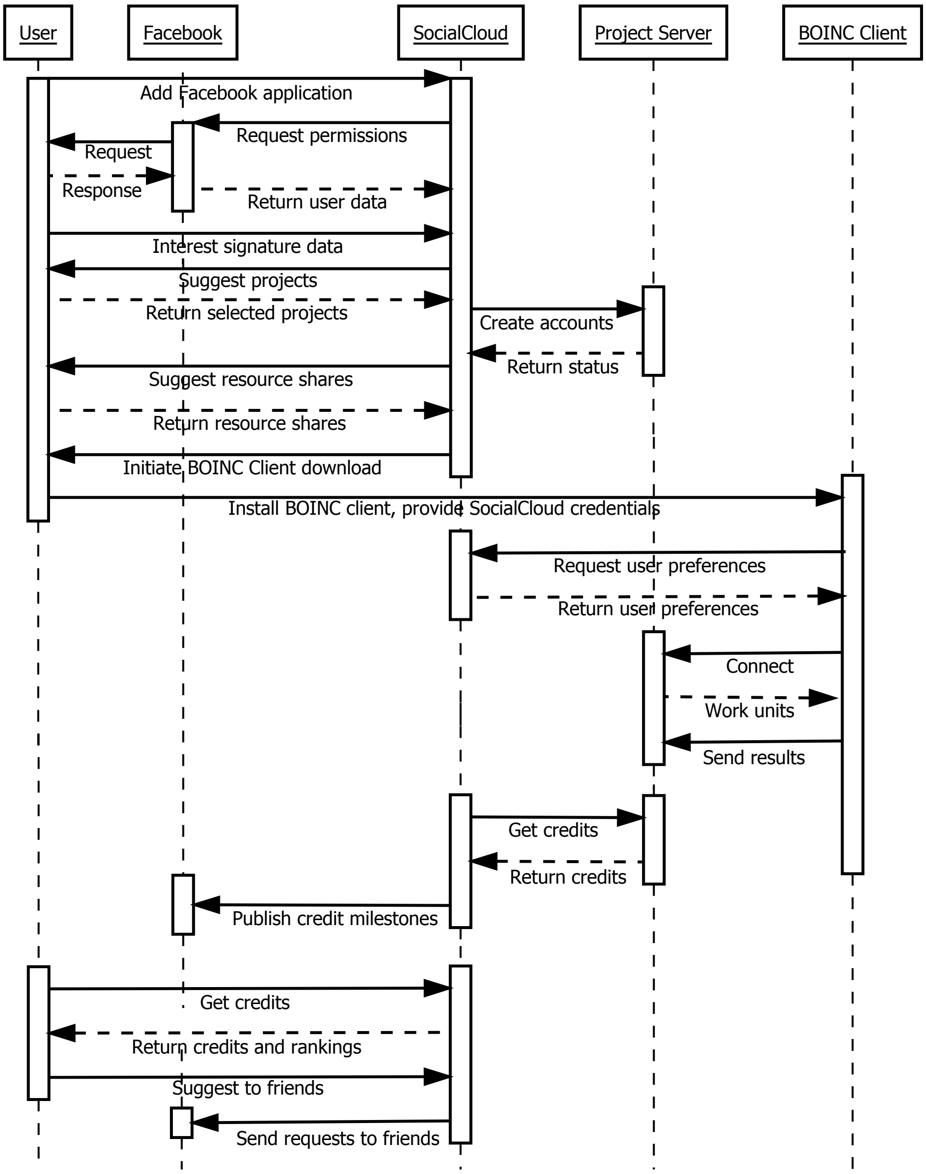

Figure 6.2: A simple example of interactions between all the actors associated with the Social Cloud for Public eResearch.

Chapter 7

Results and Analysis

In this chapter, I will go through the analysis of the contributions of the Social Cloud of Public eResearch. I will first introduce a standard dataset that is used for social network analysis and explain how simulations on it have supported several of the assertions in this thesis. I will show how bringing together social networks and volunteer computing is beneficial to the cause of the latter. Those contributions that cannot be reliably studied through simulations, due to the inability to simulate human behaviour, are supported through logical analysis – only a large scale deployment would allow conclusive statements to be made. I also share results and conclusions drawn from a limited user study that looks at effectiveness of interest/project signatures and signature distances.

7.1 Visual Analytics Benchmark

In order to study the effects of the contributions of this thesis, I used a standard social network dataset to study and perform simulations that would generate reliable and reproducible results. The dataset is part of the Social Network and Geospatial benchmark [63] and is provided by the Human-Computer Interaction Lab at the University of Maryland. It was used for the IEEE Visual Analytics Science and Technology (VAST) 2009

Challenge.

The dataset consists of two tab-delimited tables – one of which describes entities (people, cities and countries) and the other containing links between the entities.

The entities table consisted of 6016 entities with 6000 of them being persons, 12 of them being cities and the four remaining are countries. Geospatial data being of little importance for the purposes of this thesis, the 16 non-person entities were omitted. The table of connections consists of 29,888 entries of which 12 were connections between cities and countries. These 12 connections were similarly omitted.

The resultant dataset was further analysed and it was found that the minimum number of connections from any person in the set was 4. The maximum number of connections was found to be 449. The average number of connections to any given person was found to be 9. Figure 7.1 shows the distribution of friend connections in the dataset. This distribution exhibits a very small number of users having a very large number of connections while the majority of the users have a small number of connections. This distribution follows the Pareto principle or 80-20 rule [64] and is known to be representative of real world friend connections.

The graph shows that more than half of the people represented in the dataset have between 5 and 7 connections. This is a pessimistic model for simulation given that the average number of friends that a Facebook user has is around 120 [65]. So it is expected that results from real world social networks would significantly improve over the results presented in the sections following that use the VAST Challenge dataset.

7.2 Testbed

The experiment in section 7.3 was run on web server to which participants browsed through a web browser. The web application itself was written in PHP and used a MySQL database backend for data collection. The analy-

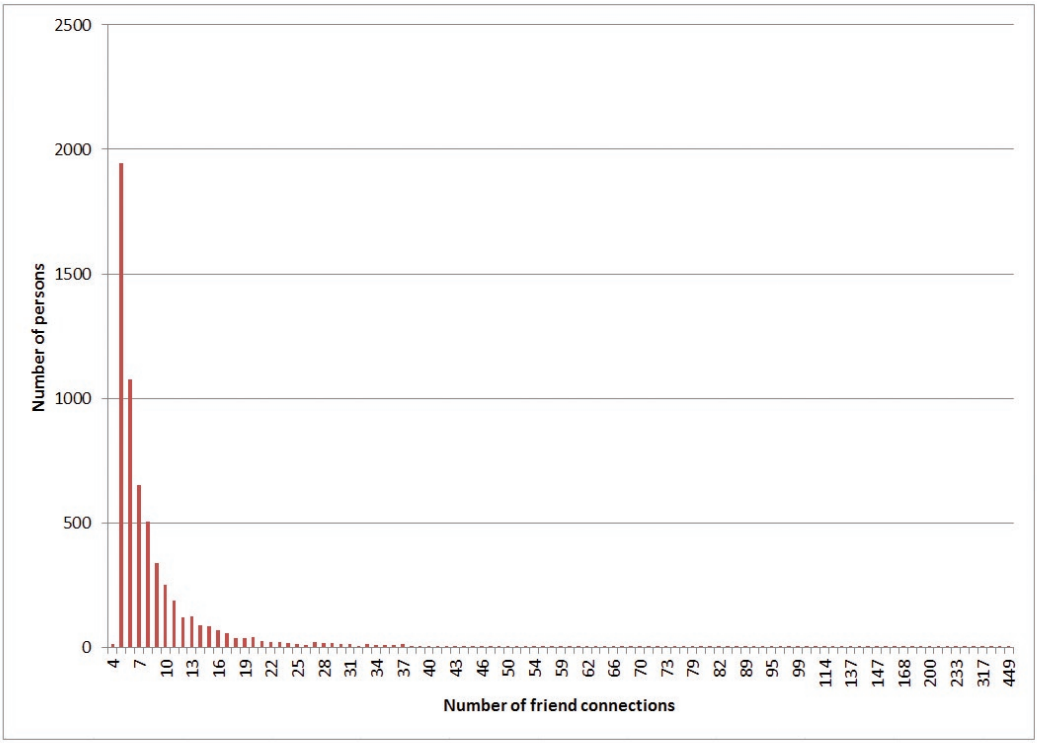

Figure 7.1: Representation of the number of connections persons in the VAST Challenge dataset have

sis of the data obtained was done using Microsoft Excel.

The simulations described in sections 7.4, 7.5 and 7.6, and the virtual environment to run them in, were written from scratch in C#. The algorithms used, described in chapter 5, were reproduced within this virtual environment. The simulations were run on a local machine. Data obtained after each simulation was ported into Microsoft Excel for in depth analysis and graphing.

7.3 Signatures and Signature Distances

In order to explore the real world benefits of the user signatures, project signatures and signature distances, I decided to perform a user study with 10 people. The users were required to provide their preferences to arrive at their user signatures, and to select 5 projects that interested them the most from a list of 27 projects. The effectiveness of the signatures and signature distances concepts would be measured by checking how many of those 5 appeared in the top 10 suggestions that would have been given by the Social Cloud.

The experiment was set up so that users described their interests on a scale of 0 to 10 each over the following interest areas:

- Astronomy
- Biology
- Chemistry
- Cryptography
- Disease
- Earth Sciences
- Game-play

- Mathematics
- Physics

The project list used in this experiment was derived from the BOINC project list at [7] and is shown in Figure 7.2. Since there was no existing user base from which to derive project signatures, I arrived at project signatures for each of those projects by carefully studying their descriptions. I also used project popularity, relative distribution of subject areas and platforms supported in arriving at final normalized project signatures.

After having their interest signatures recorded and normalized (shown in Figure 7.3), the 10 users were shown the project list along with descriptions of each and asked to select the top 5 projects that interested them. This was a blind test, and the user's interest signature was not actually used to influence the project list.

After all 10 users had completed their part of the experiment, I checked to see how many of the user's selections made their way to the top 10 of their own personalized lists. The results are given in percentages at the bottom of Figures 7.4 and 7.5.

The experiment confirmed the effectiveness of signatures and signature distances in making it easier for users to select projects that are likely to interest them.

7.4 Project Champions

In this section, I describe the experiments I ran on the VAST dataset to study the effects of introducing the concept of Project Champions (section 5.5.2.3) to the Social Cloud.

Every individual in the VAST dataset is initialized as a member of the Social Cloud for this experiment. The Social Cloud is initialized with 5 projects having 60%, 25%, 10%, 4% and 1% of the total computational time available. Each user is randomly initialized with varying resource shares

	Astronomy	Biology	Chemistry	Cryptograpl	Disease	Earth Sciences	Game-playing	Mathematics	Physics
ABC@home	0.000	0.000	0.000	0.000	0.000	0.000	0.000	45.000	0.000
CAS@home	0.000	14.211	14.211	0.000	0.000	0.000	0.000	0.000	16.579
Chess960@home	0.000	0.000	0.000	0.000	0.000	0.000	27.000	18.000	0.000
Climateprediction.net	0.000	0.000	0.000	0.000	0.000	45.000	0.000	0.000	0.000
Collatz Conjecture	0.000	0.000	0.000	0.000	0.000	0.000	0.000	45.000	0.000
DistRTgen	0.000	0.000	0.000	24.231	0.000	0.000	0.000	20.769	0.000
Docking@Home	0.000	22.500	0.000	0.000	22.500	0.000	0.000	0.000	0.000
EDGeS@Home	0.000	0.000	14.318	0.000	0.000	0.000	0.000	14.318	16.364
Einstein@Home	22.500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	22.500
Enigma@Home	0.000	0.000	0.000	24.000	0.000	0.000	0.000	21.000	0.000
GPUGrid.net	0.000	45.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ibercivis	0.000	9.265	9.265	0.000	9.265	0.000	0.000	7.941	9.265
LHC@Home	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	45.000
Malariacontrol.net	0.000	22.500	0.000	0.000	22.500	0.000	0.000	0.000	0.000
Milkyway@home	45.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
POEM@Home	0.000	25.313	0.000	0.000	19.688	0.000	0.000	0.000	0.000
PrimeGrid	0.000	0.000	0.000	23.824	0.000	0.000	0.000	21.176	0.000
QMC@Home	0.000	0.000	45.000	0.000	0.000	0.000	0.000	0.000	0.000
Rosetta@Home	0.000	23.824	0.000	0.000	21.176	0.000	0.000	0.000	0.000
SETI@Home	17.609	11.739	0.000	0.000	0.000	0.000	0.000	0.000	15.652
SIMAP	0.000	45.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Spinhenge@Home	0.000	0.000	45.000	0.000	0.000	0.000	0.000	0.000	0.000
sudoku@vtaiwan	0.000	0.000	0.000	0.000	0.000	0.000	25.313	19.688	0.000
uFluids@Home	20.769	0.000	0.000	0.000	0.000	0.000	0.000	0.000	24.231
VirtualPrairie	0.000	21.176	0.000	0.000	0.000	23.824	0.000	0.000	0.000
World Community Grid	0.000	15.577	0.000	0.000	15.577	13.846	0.000	0.000	0.000
Yoyo@home	0.000	12.857	0.000	0.000	0.000	0.000	0.000	17.143	15.000

Figure 7.2: Projects and project signatures.

RAW USER INPUT

Users	Astronom	Biology	Chemistry	Cryptogra	Disease	Earth Scie	Game-pla	Mathema	Physics
1	0	0	0	2	8	9	3	3	6
2	1	3	1	9	6	5	1	8	7
3	6	8	5	1	9	7	2	4	3
4	3	7	6	4	9	8	1	5	2
5	7	2	2	7	8	9	4	4	4
6	3	1	2	9	2	2	8	9	6
7	0	0	0	0	9	6	0	0	9
8	0	0	0	0	4	8	0	0	9
9	6	2	3	8	3	3	3	9	8
10	8	5	4	2	9	2	1	2	3

NORMALIZED VALUES

Users	Astronom	Biology	Chemistry	Cryptogra	Disease	Earth Scie	Game-pla	Mathema	Physics
1	0.000	0.000	0.000	2.903	11.613	13.065	4.355	4.355	8.710
2	1.098	3.293	1.098	9.878	6.585	5.488	1.098	8.780	7.683
3	6.000	8.000	5.000	1.000	9.000	7.000	2.000	4.000	3.000
4	3.000	7.000	6.000	4.000	9.000	8.000	1.000	5.000	2.000
5	6.702	1.915	1.915	6.702	7.660	8.617	3.830	3.830	3.830
6	3.214	1.071	2.143	9.643	2.143	2.143	8.571	9.643	6.429
7	0.000	0.000	0.000	0.000	16.875	11.250	0.000	0.000	16.875
8	0.000	0.000	0.000	0.000	8.571	17.143	0.000	0.000	19.286
9	6.000	2.000	3.000	8.000	3.000	3.000	3.000	9.000	8.000
10	10.000	6.250	5.000	2.500	11.250	2.500	1.250	2.500	3.750

Figure 7.3: Raw user inputs and normalized interest signatures.

	USER 1			USER 2			USER 3			USER 4			USER 5		
	Distances	Rank	Selections	Distances	Rank	Selections	Distances	Rank	Selections	Distances	Rank	Selections	Distances	Rank	Selections
ABC@home	45.396	21		39.451	19		44.159	25		43.128	24		44.209	21	
CAS@home	28.595	5		24.917	8		22.244	4		22.977	3		26.777	6	
Chess960@home	32.997	15		31.611	18		32.955	16		33.226	18		31.343	18	
Climateprediction.net	35.734	19	Yes	43.043	22		40.988	21	Yes	39.875	19	Yes	39.033	19	Yes
Collatz Conjecture	45.396	21		39.451	19		44.159	25		43.128	24		44.209	21	
DistRTgen	33.536	18		22.287	6		32.998	18		30.033	14		28.438	10	
Docking@Home	30.293	8		29.827	10		23.076	5	Yes	24.031	5	Yes	29.253	13	
EDGeS@Home	26.378	3	Yes	21.567	3	Yes	24.635	10		24.132	7		25.730	4	
Einstein@Home	32.378	13		30.643	16		29.875	12		32.749	16		28.507	11	
Enigma@Home	33.504	17		22.265	5	Yes	32.955	16		30.000	13		28.435	8	
GPUGrid.net	49.525	23		45.280	23		39.875	19		40.988	20		46.117	24	
Ibercivis	19.709	2		15.584	1		12.831	1		12.934	1		18.325	1	
LHC@Home	40.852	20		40.684	21		45.166	27		46.152	27		44.209	21	
Malariacontrol.net	30.293	8	Yes	29.827	10		23.076	5	Yes	24.031	5	Yes	29.253	13	Yes
Milkyway@home	49.525	23		47.412	27		42.071	22		45.166	26		41.181	20	
POEM@Home	31.604	11		30.397	14		23.536	9	Yes	24.588	9		30.065	16	
PrimeGrid	33.481	16		22.251	4	Yes	32.923	15		29.978	12		28.435	9	Yes
QMC@Home	49.525	23		47.412	25		43.128	23		42.071	22		46.117	24	
Rosetta@Home	30.853	10		30.031	12		23.209	8	Yes	24.214	8	Yes	29.571	15	
SETI@Home	29.120	7	Yes	25.652	9	Yes	22.019	3		25.402	11		23.792	3	Yes
SIMAP	49.525	23		45.280	23		39.875	19		40.988	20		46.117	24	
Spinhenge@Home	49.525	23		47.412	25		43.128	23		42.071	22		46.117	24	
sudoku@vtaiwan	32.621	14		30.800	17		32.474	14		32.647	15		30.947	17	Yes
uFluids@Home	32.003	12		30.368	13		30.148	13		32.893	17		28.785	12	
VirtualPrairie	28.658	6		30.606	15		25.073	11		24.967	10		28.256	7	
World Community Grid	19.524	1	Yes	23.215	7	Yes	15.445	2		15.556	2	Yes	20.352	2	
Yoyo@home	26.483	4		19.739	2		23.159	7		23.617	4		25.746	5	
			80%			100%			80%			80%			40%

Figure 7.4: Results for Users 1 to 5.

	USER 6			USER 7			USER 8			USER 9			USER 10		
	Distances	Rank	Selections	Distances	Rank	Selections	Distances	Rank	Selections	Distances	Rank	Selections	Distances	Rank	Selections
ABC@home	38.512	19		52.164	21		52.576	21		38.730	19		46.098	25	
CAS@home	26.462	10		28.553	3		27.903	3		23.588	8		23.657	6	
Chess960@home	23.854	8	Yes	41.822	20		42.335	20	Yes	29.189	13		34.936	18	
Climateprediction.net	46.456	23		41.335	15	Yes	34.949	11		45.166	23		46.098	25	
Collatz Conjecture	38.512	19	Yes	52.164	21		52.576	21		38.730	19		46.098	25	
DistRTgen	21.832	6		41.408	18		41.926	18		23.280	7		33.444	16	
Docking@Home	34.553	14		30.809	8		36.960	12		32.749	14		23.452	4	
EDGeS@Home	21.352	2		28.664	4		28.034	4		18.903	2	Yes	25.793	10	
Einstein@Home	30.076	13		30.809	8	Yes	29.731	8	Yes	25.836	10		26.810	11	
Enigma@Home	21.798	5	Yes	41.390	17		41.908	17		23.238	6	Yes	33.422	15	
GPUGrid.net	47.483	26		52.164	21		52.576	21		46.152	26		42.279	20	
Ibercivis	18.997	1		21.843	1		25.090	1		15.874	1		14.328	1	
LHC@Home	42.100	21		34.675	14		32.071	10	Yes	39.875	21		44.861	24	
Malariacontrol.net	34.553	14		30.809	8	Yes	36.960	12		32.749	14		23.452	4	Yes
Milkyway@home	45.406	22		52.164	21		52.576	21		42.071	22		38.079	19	
POEM@Home	34.867	18		32.557	12		37.817	15		33.075	18		24.371	8	Yes
PrimeGrid	21.775	4	Yes	41.377	16		41.896	16		23.209	5	Yes	33.407	14	
QMC@Home	46.456	23		52.164	21		52.576	21		45.166	23		43.589	22	
Rosetta@Home	34.644	17		31.582	11		37.313	14		32.843	17		23.806	7	Yes
SETI@Home	26.063	9		29.338	5	Yes	28.782	6		21.660	4	Yes	20.039	3	Yes
SIMAP	47.483	26		52.164	21		52.576	21		46.152	26		42.279	20	
Spinhenge@Home	46.456	23		52.164	21		52.576	21		45.166	23		43.589	22	
sudoku@vtaiwan	23.253	7		41.526	19		42.043	19		28.409	12	Yes	34.520	17	
uFluids@Home	29.991	12		29.947	7		28.691	5	Yes	25.818	9		27.320	12	
VirtualPrairie	34.562	16		34.294	13		30.635	9		32.762	16		30.938	13	
World Community Grid	29.044	11		23.148	2	Yes	25.972	2	Yes	26.891	11		19.687	2	Yes
Yoyo@home	21.429	3	Yes	29.564	6		29.067	7		19.214	3		25.500	9	
			80%			80%			80%			80%			100%

Figure 7.5: Results for Users 6 to 10.

totalling 100 such that the overall percentages in the Social Cloud are as described. These percentages were selected to reflect the inequitable distribution of resources amongst BOINC projects currently.

The main goal of the Project Champion incentive is allow lesser known projects to obtain more computational resources. When Project Champion incentives are enabled, users try to become Project Champions of as many projects as possible within their social cloud.

Without Project Champion incentives, the initial resource share percentages would not change by much – as BOINC Projects have historically seen.

7.4.1 User Behaviour

There were two things that needed to be modelled to realistically represent user behaviour for Compute Magnate simulations:

- Participation – Users changes their project selections once a week if required to maximize their chances of being Project Champions. The probability of them participating in a particular week is inversely proportional to a *disinterest factor* that is assigned to them from a Poisson distribution.

- Share Redistribution - Users select projects requiring the least shares (in ascending order) to bring them the project champion status within their social cloud and reallocate shares accordingly.

7.4.2 The Simulations

The experiments were run for values of λ for the Poisson distribution set to 1, 3, 5 and 10, and the results can be seen in Figures 7.6, 7.7, 7.8 and 7.9 respectively. The values from the Poisson distributions determined the *disinterest factor* attributed to each user. All the experiments were run for a simulated duration of 52 weeks (1 year). If Project Champion incentives worked, we would be seeing a redistribution of resources among projects to more equitable levels.

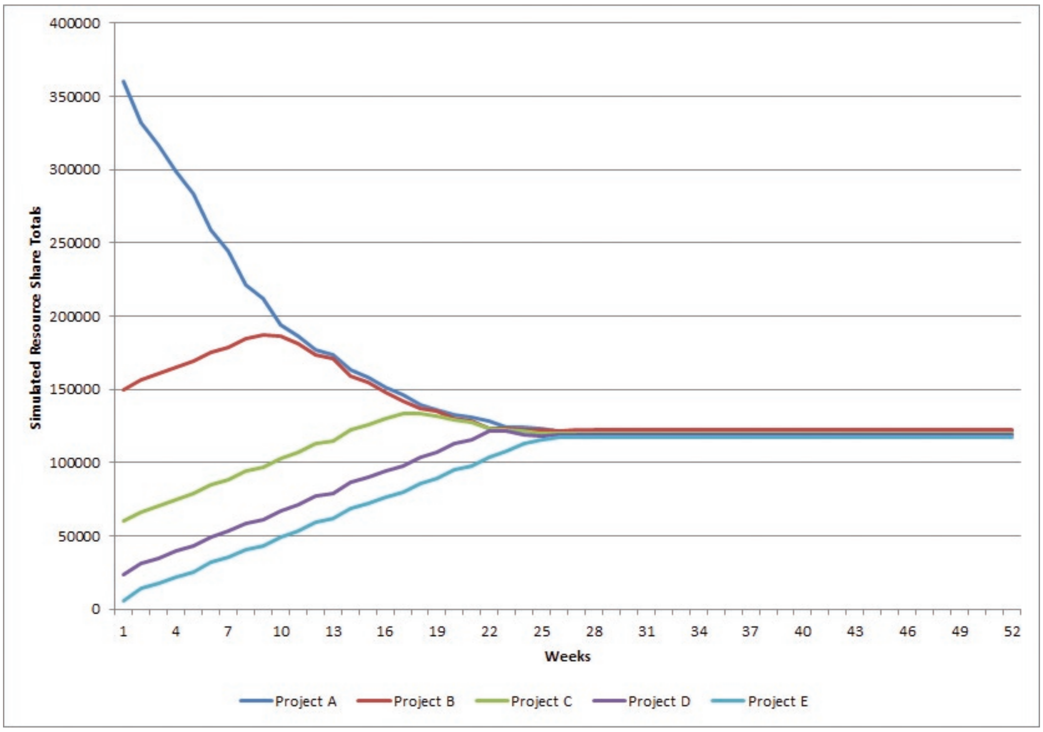

Figure 7.6: Project Champion Simulations – Representation of project resource share distributions over a simulated duration of 1 year. The *disinterest factor* of every user was obtained from a Poisson distribution of λ **value = 1**. What is seen here is the global distribution of computational time among projects equalizing as users try to become Project Champions by prioritizing projects with the least support in their social cloud. The disinterest factor governs a given user's inclination to participate in this process.

The Project Champion incentive mechanism is very sensitive to user participation. Low levels of user participation can significantly reduce the effects of Project Champion incentives. However, as can be seen from the results, the Project Champion incentives are quite capable of producing

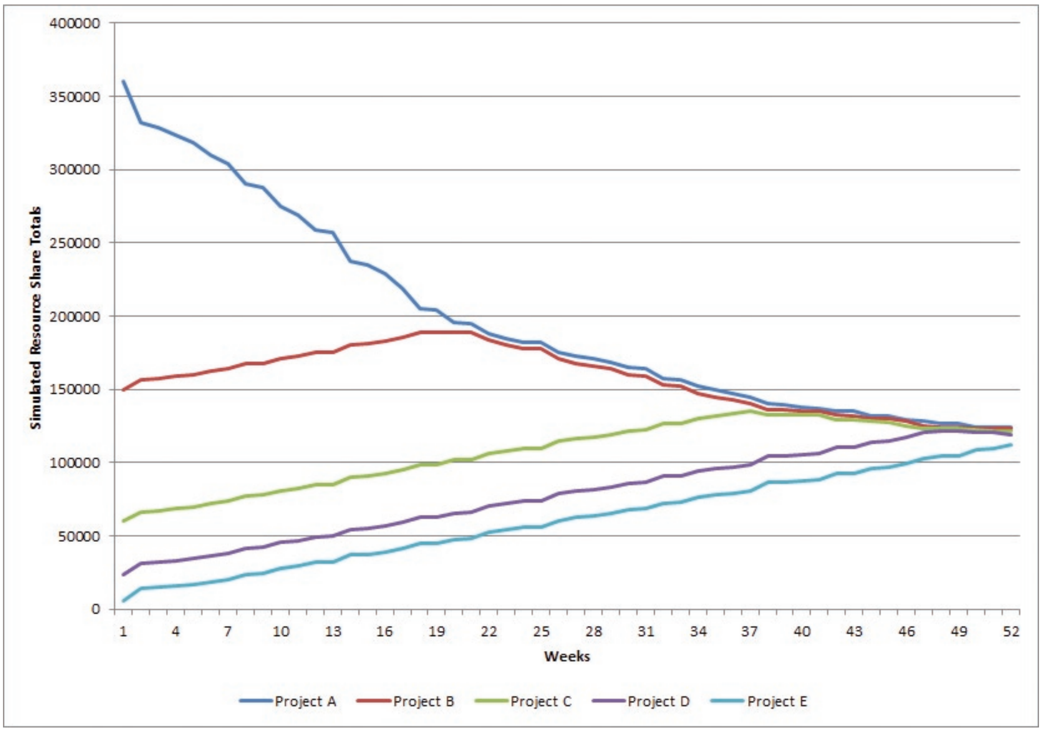

Figure 7.7: Project Champion Simulations – Representation of project resource share distributions over a simulated duration of 1 year. The *disinterest factor* of every user was obtained from a Poisson distribution of λ **value = 3**. What is seen here is the global distribution of computational time among projects equalizing as users try to become Project Champions by prioritizing projects with the least support in their social cloud. The effect of slightly higher disinterest factors are visible in this graph.

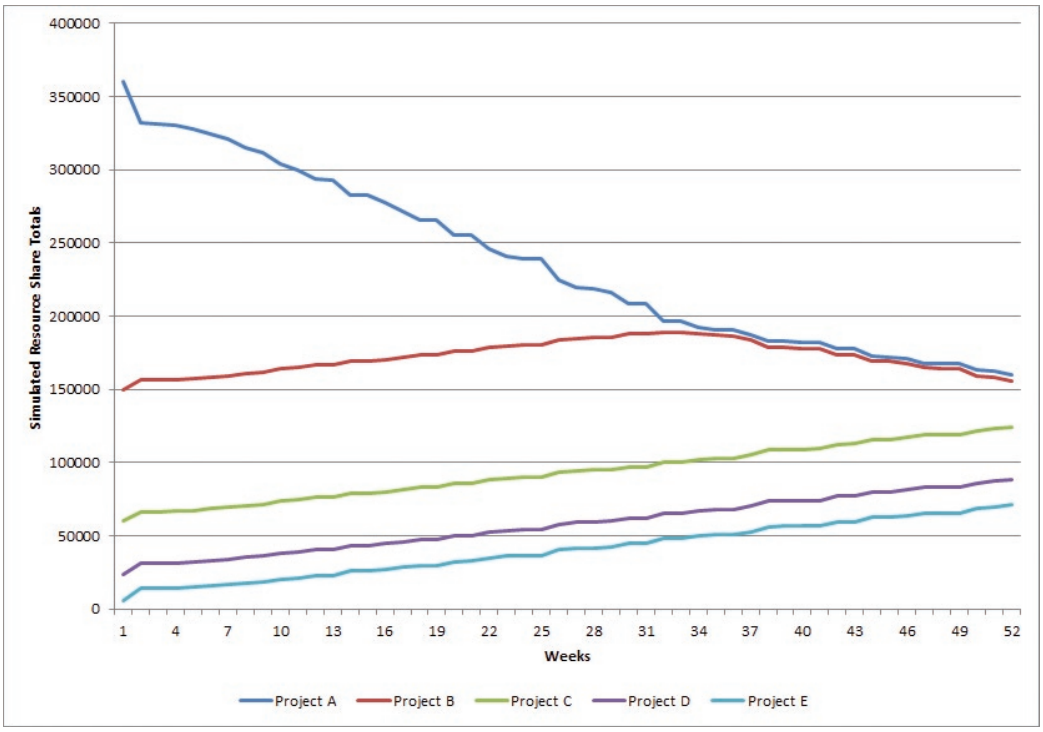

Figure 7.8: Project Champion Simulations – Representation of project resource share distributions over a simulated duration of 1 year. The *disinterest factor* of every user was obtained from a Poisson distribution of λ **value = 5**. What is seen here is the global distribution of computational time among projects equalizing as users try to become Project Champions by prioritizing projects with the least support in their social cloud. The effect of higher disinterest factors are easily apparent here.

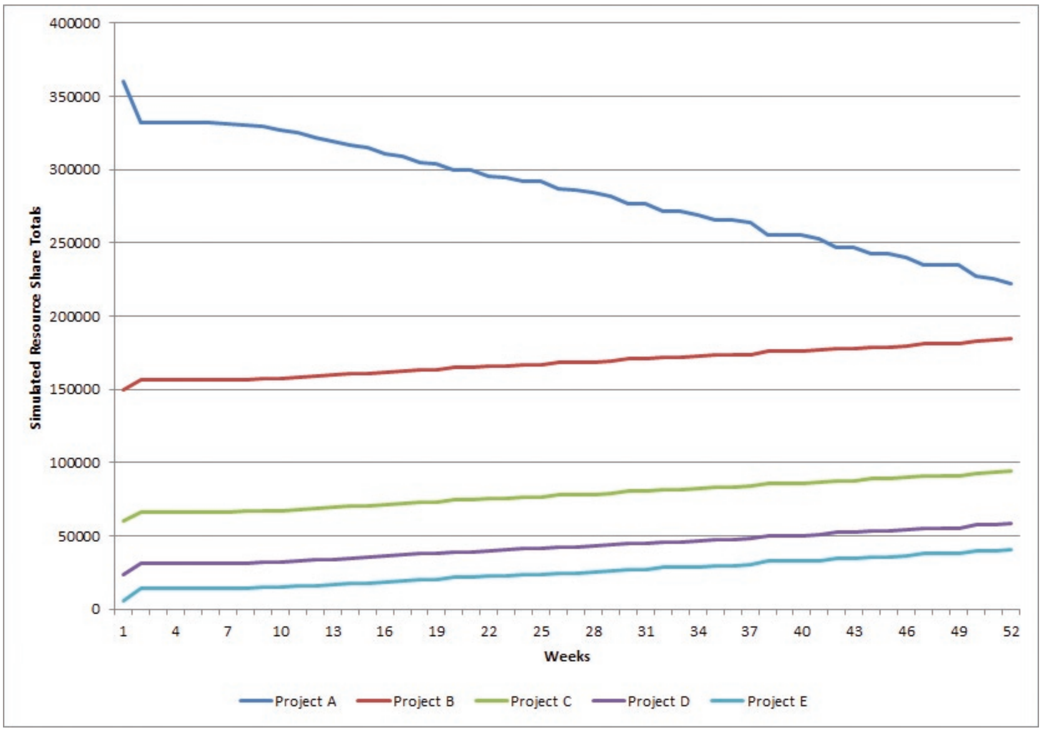

Figure 7.9: Project Champion Simulations – Representation of project resource share distributions over a simulated duration of 1 year. The *disinterest factor* of every user was obtained from a Poisson distribution of λ **value = 10**. Even when the disinterest factors are set this high, the Project Champion incentive has the desired effect albeit in a damped manner.

the desired outcome.

7.5 Social Anchors

In this section, I describe the experiments I ran on the VAST dataset to study the effects of introducing the concept of Social Anchors (section 5.5.2.2) to the Social Cloud.

At all times, the entire set of social network users fall into one of two categories - users that are not part of the Social Cloud and users that are part of the Social Cloud. All experiments in this section have the initial seed state where 15 random members of social network start off as members of the Social Cloud also. The selection of the initial 15 was found to have no statistical significance to the final outcome of the experiments. The number 15 was selected as it is the same percentage (0.25%) of the total VAST social network size (6000) as is the current number of BOINC users (2.2 million [4]) to the current size of the Facebook user base (800 million [5]).

All the experiments in this section were run with and without the use of the social anchor incentives. The idea being this approach being that if the concept of Social Anchor and the incentives behind it worked, we would be seeing a faster growth of the Social Cloud when it is used. This is clearly the case as can be seen in figures 7.10, 7.11, 7.12, 7.13 and 7.14.

7.5.1 User Behaviour

I modelled user behaviour over various key aspects to ensure that the experiments were as close to what we would see in the real world as possible.

For the average user in the social network to join the Social Cloud, they must be asked to do so by one of their friends. It would be disingenuous to assume to that all users would join the Social Cloud after just a single request from their friends, or that even every user in the social network would join if asked several times. There may be a few users would join on a single request, some might require more than that and some may require to be asked so many times that it can be assumed that they would never join.

Being unable to find any single deterministic way to set the number of requests that would be required for the average user to join, I decided to use Poisson distributions with a range of different values of λ (with and without Social Anchor incentives) to conclusively determine the effect of Social Anchor incentives on the rate of growth of the Social Cloud.

So in the experiments, every user is initialized with the required number of requests from a Poisson distribution. Every time the user receives a request, the requests required to join is reduced by one. Once the requests required wind down to zero, the user joins the Social Cloud.

Every user that joins sends out a one time request to join to all their friends. Following that, once a week, every user in the Social Cloud requests one of their friends (that is not in the Social Cloud) to join the Social Cloud. When Social Anchor incentives are not being applied, the user selects an eligible friend at random. When Social Anchor incentives are applied, the user selects a friend with the highest social value among all their friends not part of the Social Cloud yet.

7.5.2 The Simulations

The experiments were run for values of λ for the Poisson distribution set to 3, 5, 7, 10 and 15 - both with and without Social Anchor incentives. The values from the Poisson distributions determined the number of requests a user would have to receive before they decide to give in and join the Social Cloud. All the experiments were run for a simulated duration of 52 weeks (1 year).

The results from the simulations show that irrespective of the nature of the response to the Social Cloud for Public eResearch by users, the Social

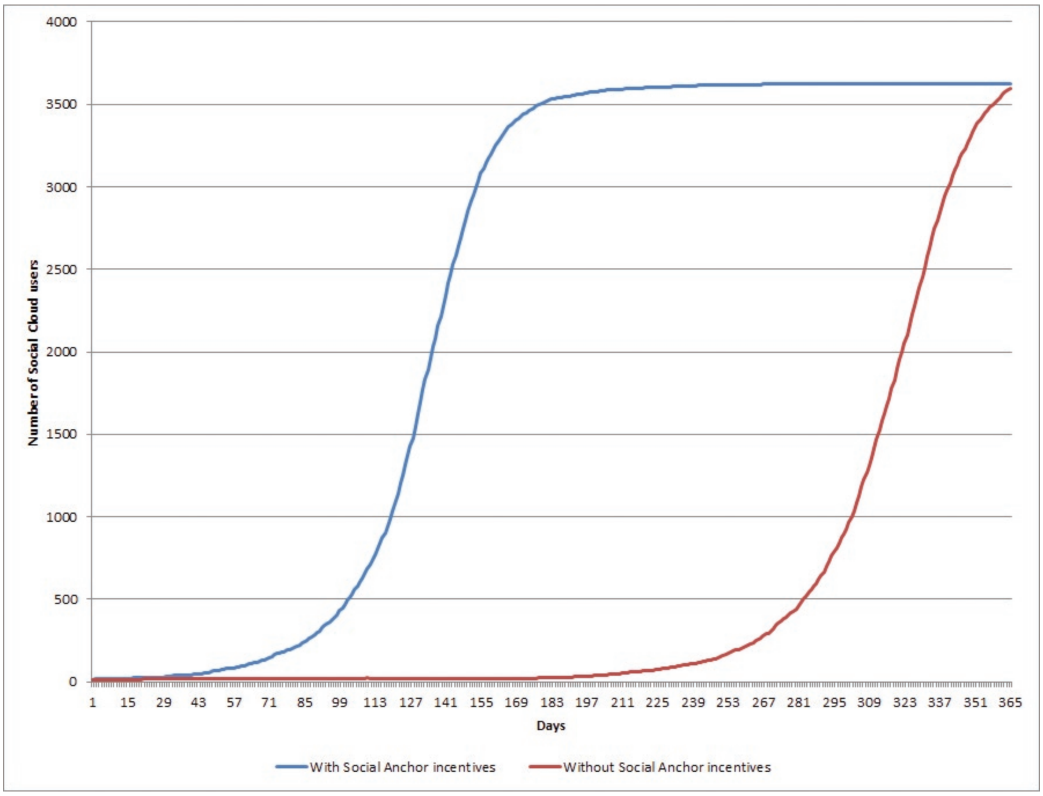

Figure 7.10: Social Anchor Simulations – Growth of the Social Cloud over a simulated duration of 1 year. The number of requests each member of the social network had to receive before they decided to become a member of the Social Cloud was obtained from a Poisson distribution of λ **value = 3**. It is seen that the Social Cloud with Social Anchor incentives hits a steep growth curve almost 6 months earlier than if those incentives were not present. The growth eventually stops as users requiring to be asked to join a very large number of times are unlikely to have been asked that many times.

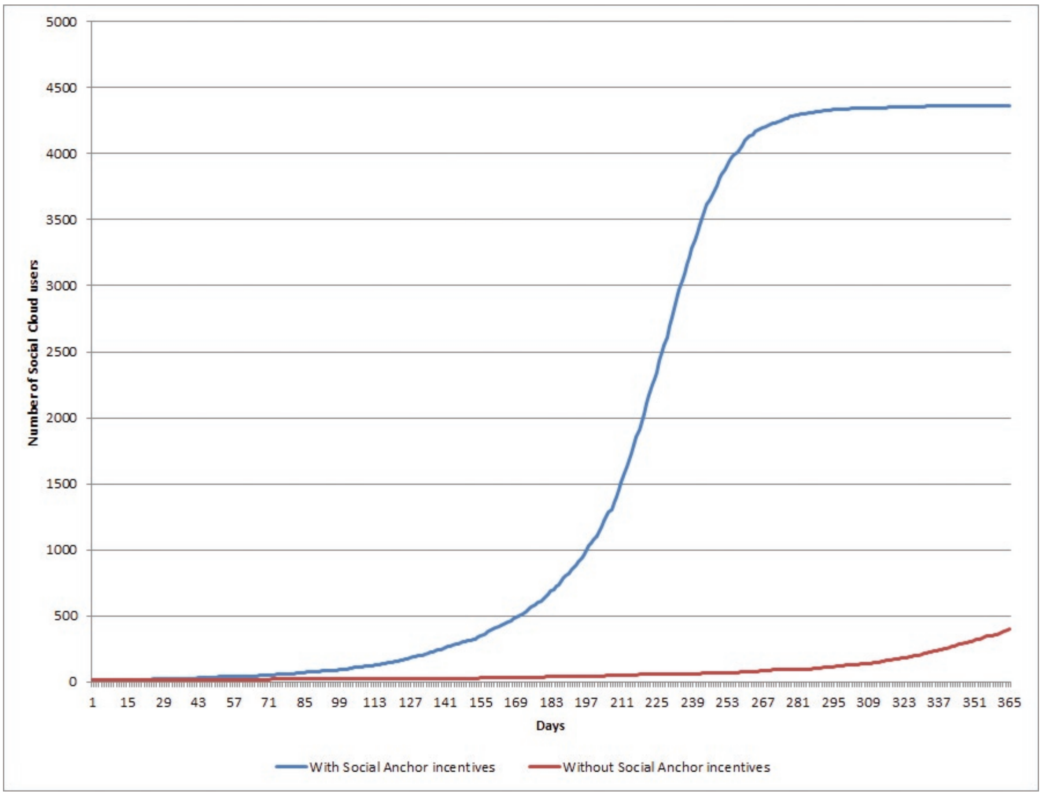

Figure 7.11: Social Anchor Simulations – Growth of the Social Cloud over a simulated duration of 1 year. The number of requests each member of the social network had to receive before they decided to become a member of the Social Cloud was obtained from a Poisson distribution of λ **value = 5**. Despite the higher resistance from social network users to join the Social Cloud, the Social Cloud still manages to hit the *viral* growth curve within a year.

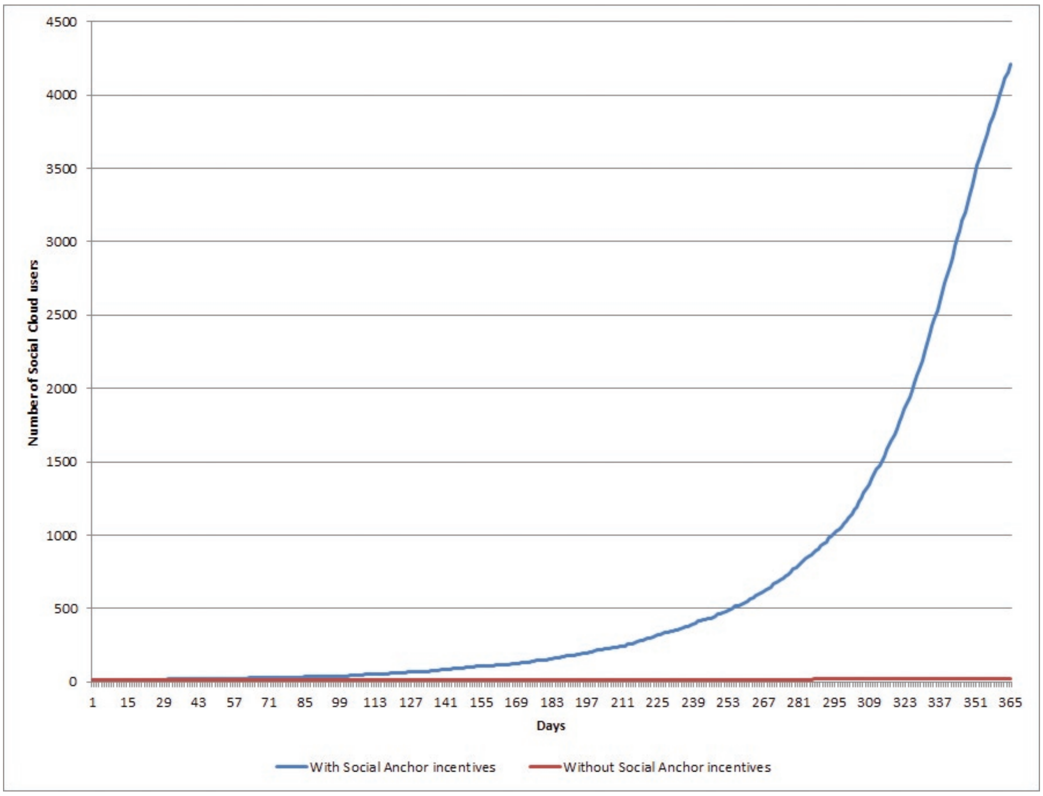

Figure 7.12: Social Anchor Simulations – Growth of the Social Cloud over a simulated duration of 1 year. The number of requests each member of the social network had to receive before they decided to become a member of the Social Cloud was obtained from a Poisson distribution of λ **value = 7**. The resistance from users is sufficiently high for viral growth to start presenting itself towards the end of the simulated year.

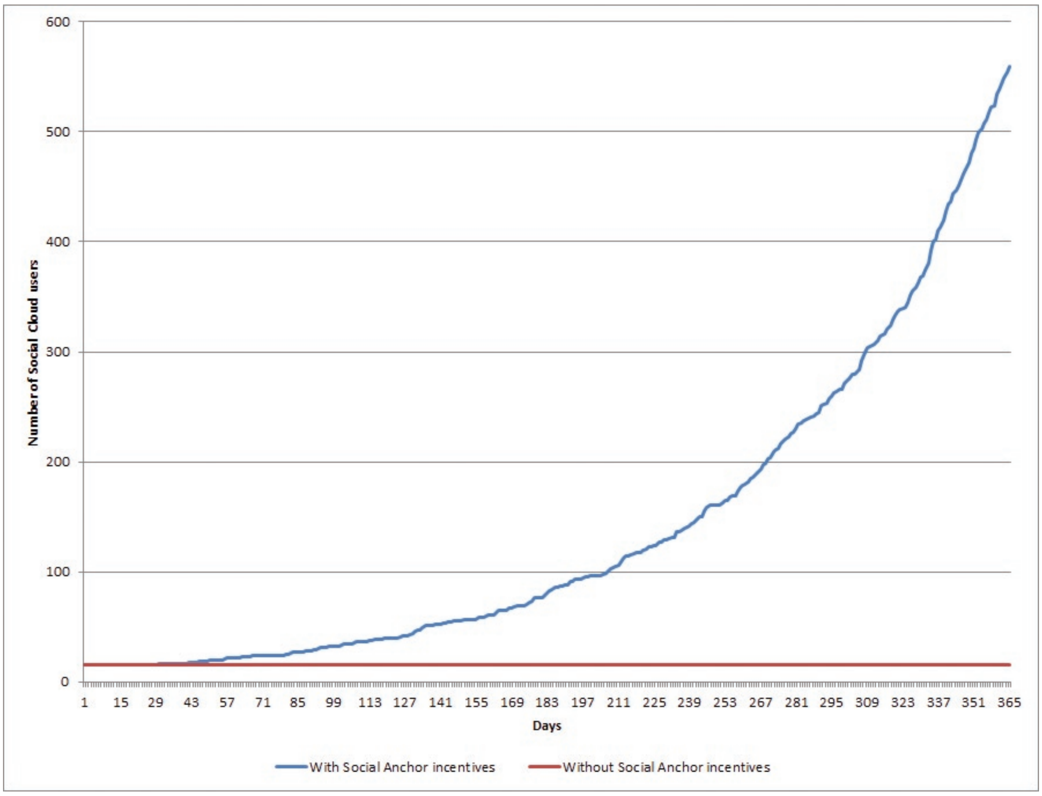

Figure 7.13: Social Anchor Simulations – Growth of the Social Cloud over a simulated duration of 1 year. The number of requests each member of the social network had to receive before they decided to become a member of the Social Cloud was obtained from a Poisson distribution of λ **value = 10**. The growth of the Social Cloud is quite subdued here in absolute terms despite the slope of the graph.

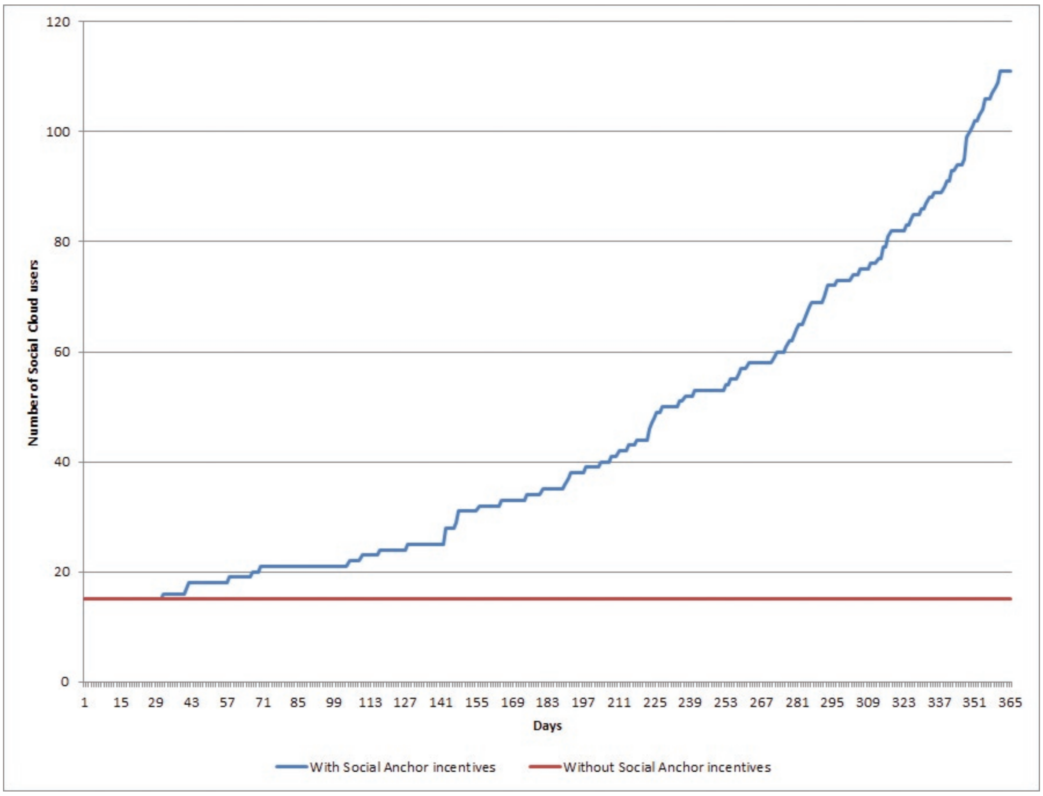

Figure 7.14: Social Anchor Simulations – Growth of the Social Cloud over a simulated duration of 1 year. The number of requests each member of the social network had to receive before they decided to become a member of the Social Cloud was obtained from a Poisson distribution of λ **value = 15**. Even with unrealistically high levels of simulated resistance, the Social Cloud manages to see growth with Social Anchor incentives which the non-incentivised cloud sees no growth at all for the entire year.

Cloud grows much faster with the Social Anchor incentives in place than without. The only effect of the responsivity is a slower growth but it affects both situations (with and without Social Anchor incentives) similarly.

7.6 Compute Magnates

In this section, I describe the experiments I ran on the VAST dataset to study the effects of introducing the concept of Compute Magnates (section 5.5.2.3) to the Social Cloud.

Every individual in the VAST dataset is initialized as a member of the Social Cloud for this experiment. Every user's contribution level is represented as a percentage of the maximum possible contributions they could have made in the given time period. In the ideal scenario, every user would be contributing at 100% – having their system running all the time.

The goal of the Compute Magnate incentive is to raise the combined contributions of all users in the Social Cloud as high as possible sustainably. When Compute Magnate incentives are enabled, active users (conservatively deemed as those with $> 25\%$ contribution levels) approach between 1 and 3 friends in the Social Cloud every week and request them to contribute more (to raise their own compute scores).

Without compute magnates incentives, what is seen is only around 12% of volunteers running BOINC actively contribute [4] at any given time.

7.6.1 User Behaviour

There were several things that needed to be modelled to realistically represent user behaviour for Compute Magnate simulations:

- Decay – The user's propensity to lose interest in contributing over time is represented as a *disinterest factor*. The disinterest factor was used to decay the contributions of every user in the Social Cloud over time – this was used to reflect a natural fall in user interest over time in the absence of external stimuli. The disinterest factor was obtained from Poisson distributions with varying values of λ.

- Participation – The user's propensity to ensure that their friends were not losing interest in contributing was considered. This was directly linked to their level of contributions at the relevant point in time as well as their disinterest factor with respect to the λ value. A user that is contributing more at a point in time is considered to be more likely to ensure more of their friends are contributing at that point in time. Users with a disinterest factor value greater than λ were set to never approach their friends. Selecting which friends to approach in a given week was done based on compute value calculations described in chapter 5.

- Growth – The magnitude of the jump in contributions when a user is requested by a friend to contribute more was inversely proportional to their (the user's) disinterest factor and directly proportional to the difference between their contributions at that moment in time and the maximum (100%). So a user with a high disinterest factor does not improve their contributions as much as another with a lower disinterest factor.

7.6.2 The Simulations

The experiments were run for values of λ for the Poisson distribution set to 5, 10, 15 and 20, and the results can be seen in Figures 7.15, 7.17, 7.19 and 7.21 respectively. The values from the Poisson distributions determined the *disinterest factor* attributed to each user. All the experiments were run for a simulated duration of 52 weeks (1 year). The initial combined contribution level was set to 12% which is representative of what BOINC currently sees [4] – assuming that all active users are contributing as much as they can, as more detailed data is not available.

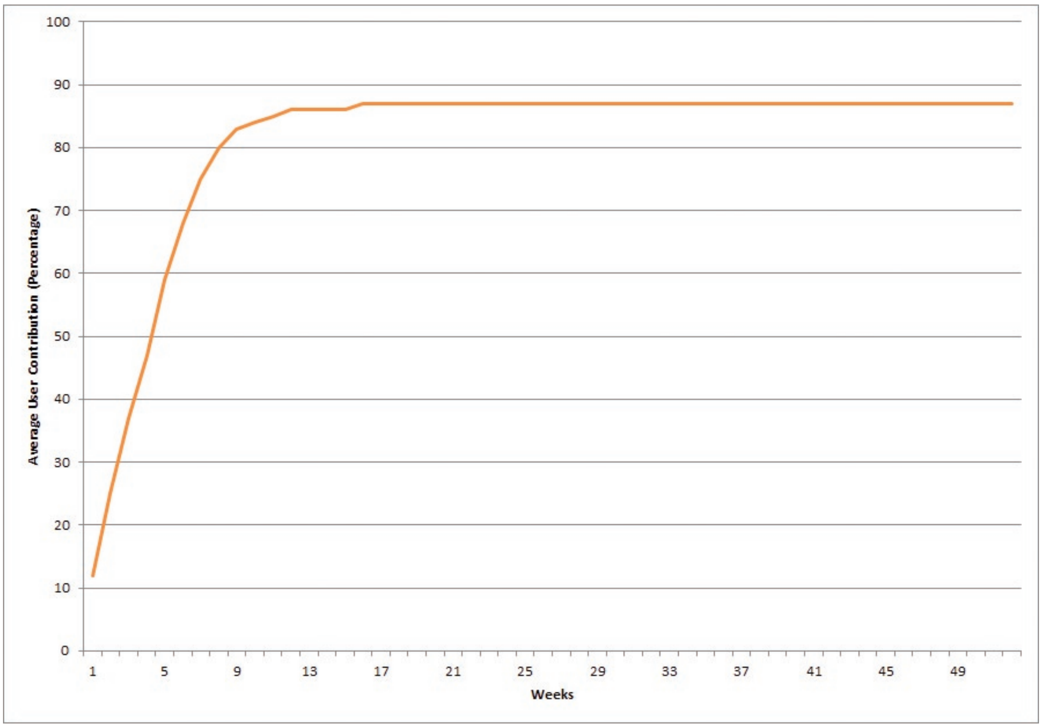

Figure 7.15: Compute Magnate Simulations – Increase in user contributions in a year with compute magnate incentives. The *disinterest factor* of every user was obtained from a Poisson distribution of λ **value = 5**. The steep growth is effected by users applying pressure on their friends to contribute more.

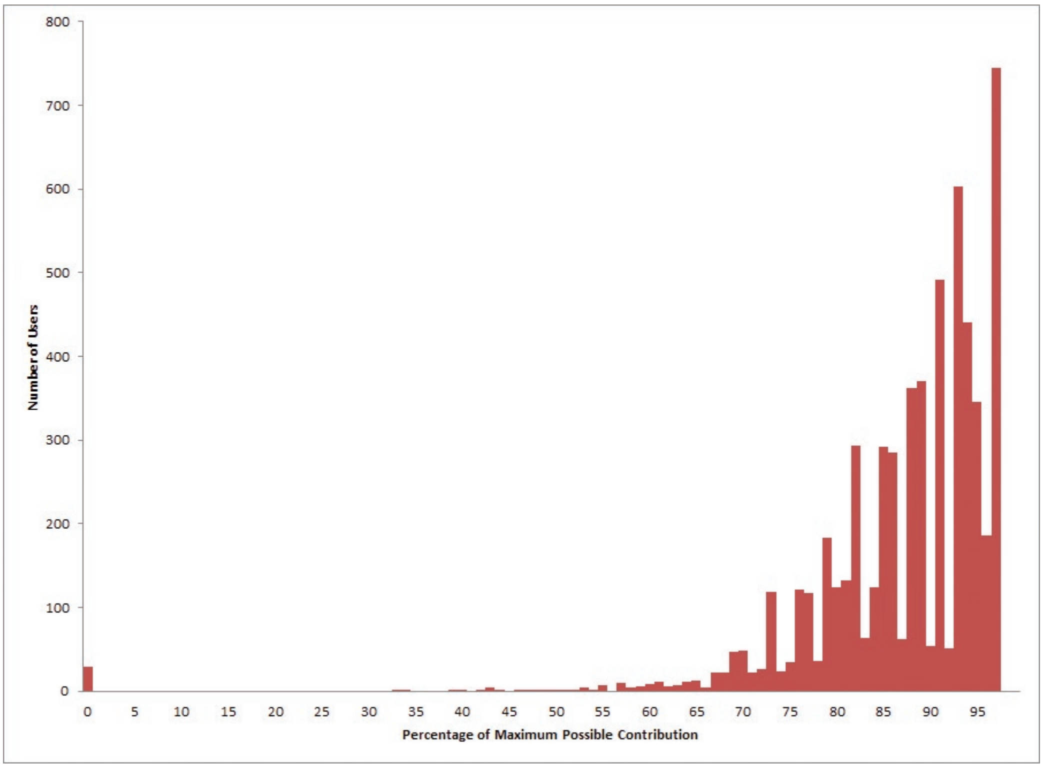

Figure 7.16: Compute Magnate Simulations – The distribution of the contribution levels of all users at the end of a year (disinterest factors obtained from a Poisson distribution of λ = **5**). What we see here is that most of the users are contributing as much as they can (between 80% and 100% of their maximums). There are a few users that lost interest and stopped contributing altogether.

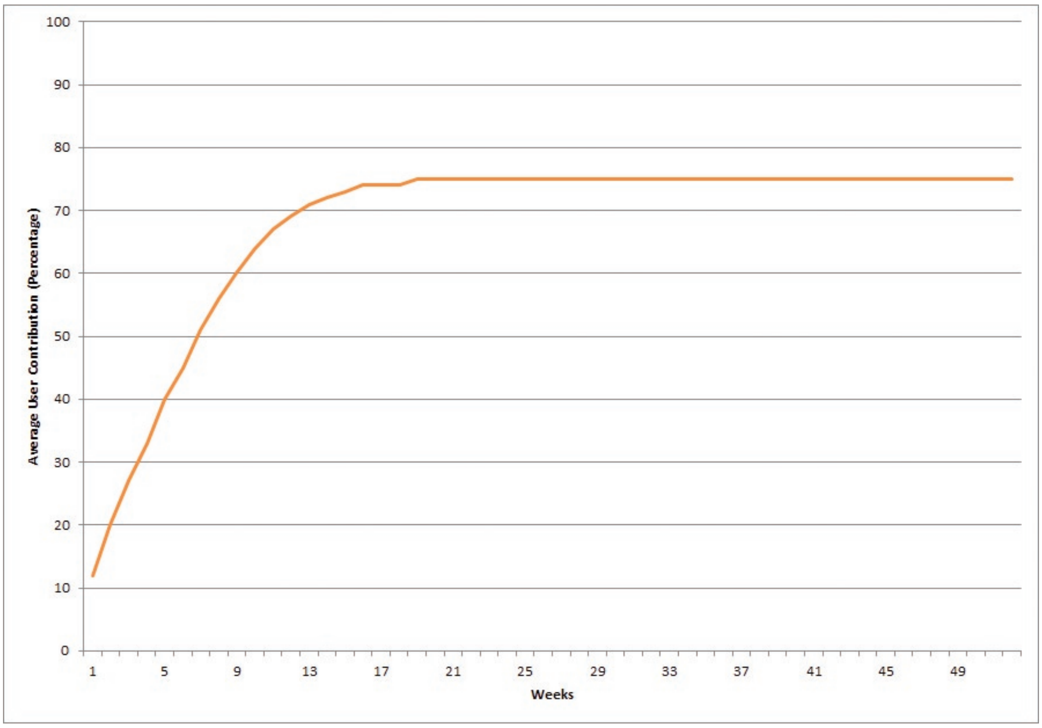

Figure 7.17: Compute Magnate Simulations – Increase in user contributions in a year with compute magnate incentives. The *disinterest factor* of every user was obtained from a Poisson distribution of λ **value = 10**. The lower plateau of the curve is due to the increased resistance from users and due to some users dropping out.

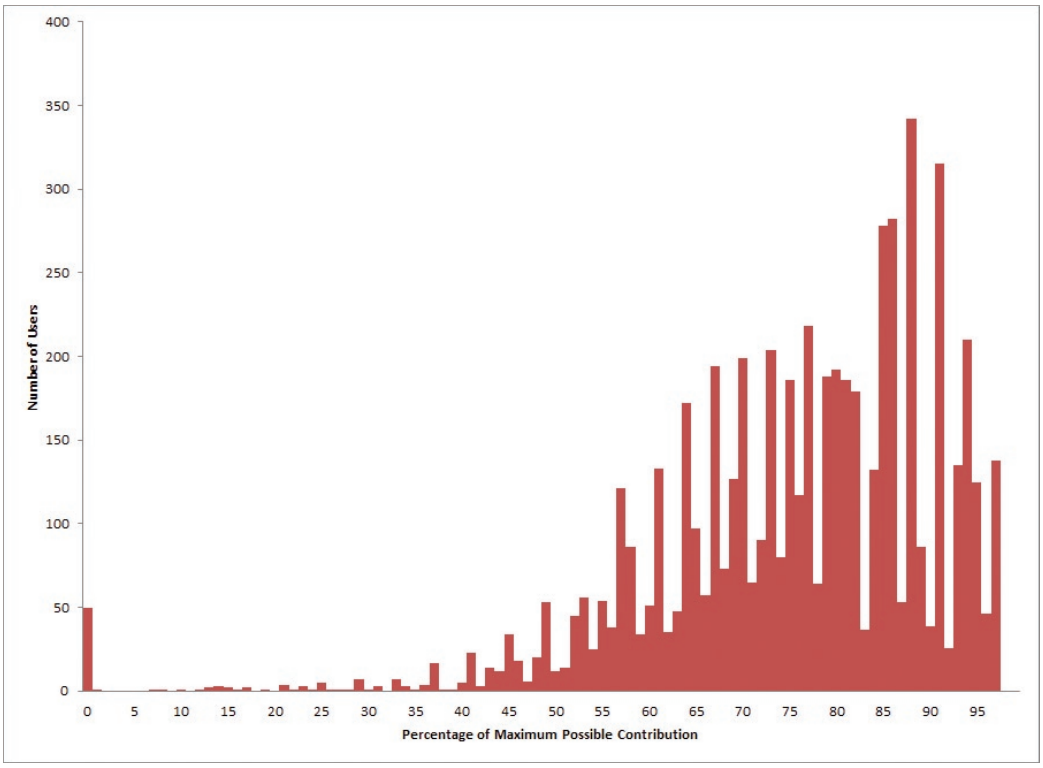

Figure 7.18: Compute Magnate Simulations – The distribution of the contribution levels of all users at the end of a year (disinterest factors obtained from a Poisson distribution of λ **= 10**). What we see here is thatthere is some resistance amongst the users to contributing as much as they can (most are doing between 60% and 95% of their maximums). There are about 50 users that lost interest and stopped contrbuting.

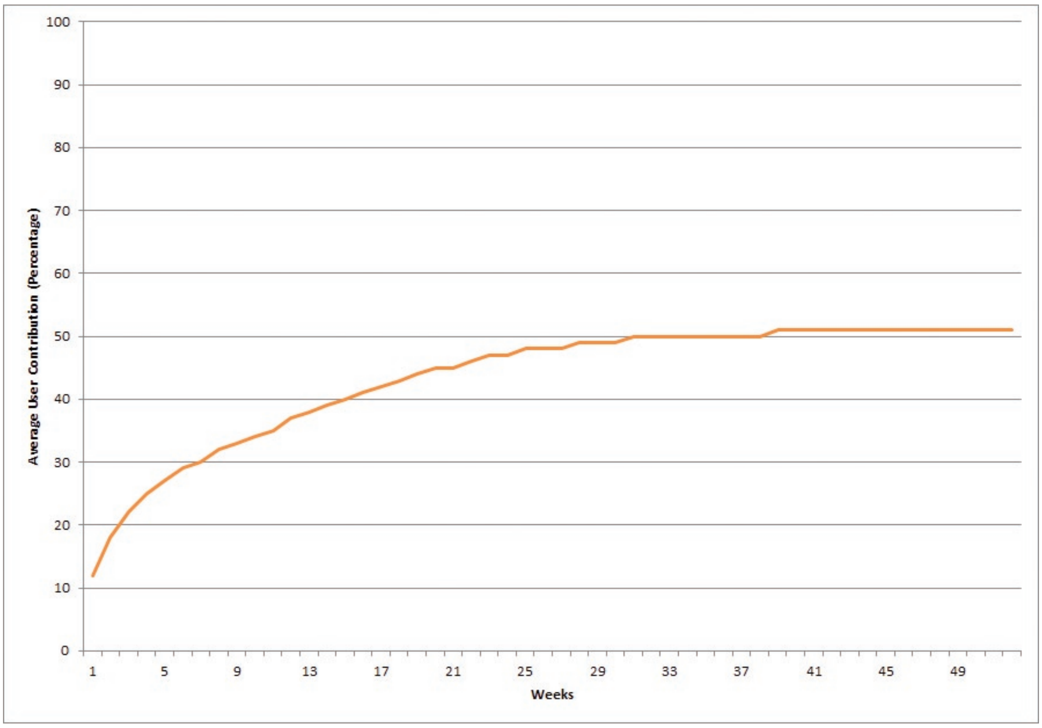

Figure 7.19: Compute Magnate Simulations – Increase in user contributions in a year with compute magnate incentives. The *disinterest factor* of every user was obtained from a Poisson distribution of λ **value = 15**. Here we see that despite the high levels of resistance, the overall levels of contribution still rises.

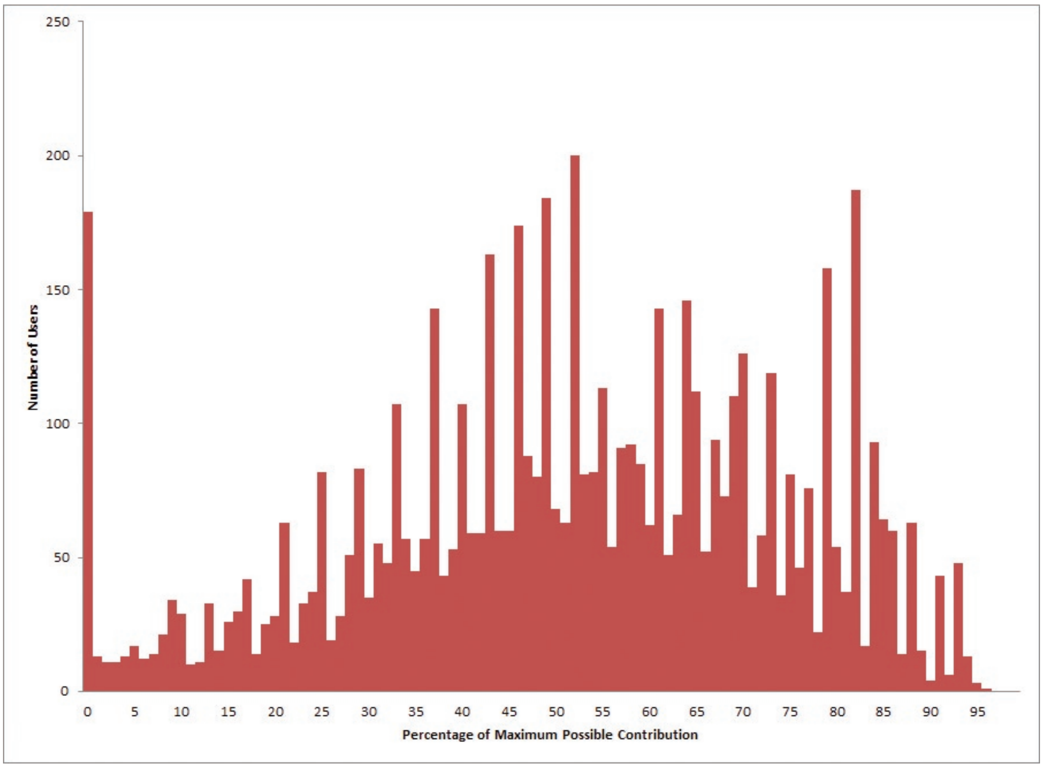

Figure 7.20: Compute Magnate Simulations – The distribution of the contribution levels of all users at the end of a year (disinterest factors obtained from a Poisson distribution of λ = **15**). There is significant resistance seen here amongst users to contributing as much as they can. There are a large number of users that are not contributing actively anymore at this high simulated resistance level.

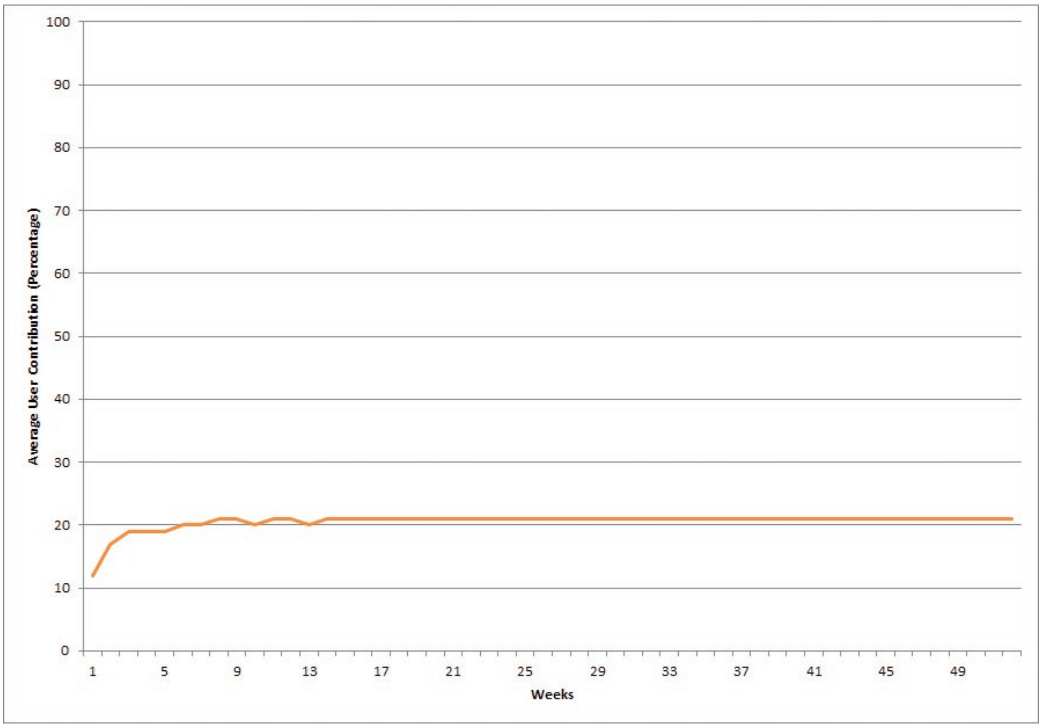

Figure 7.21: Compute Magnate Simulations – Increase in user contributions in a year with compute magnate incentives. The *disinterest factor* of every user was obtained from a Poisson distribution of λ **value = 20**. The growth in contributions is far more modest here but it is significant nonetheless.

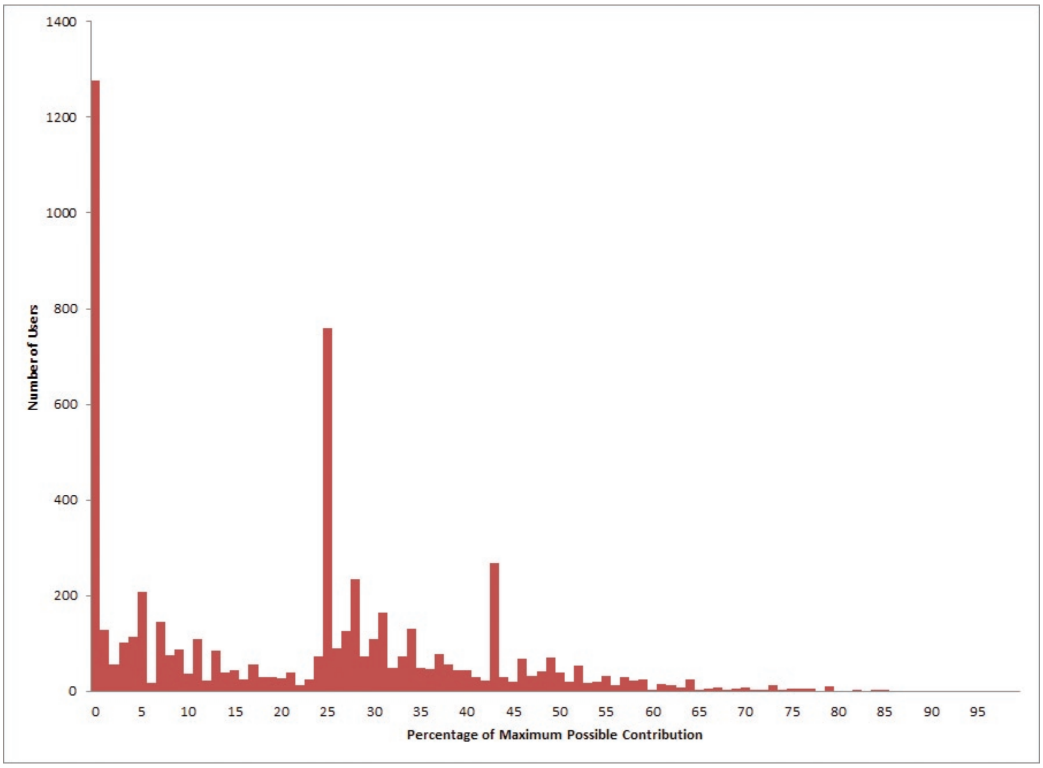

Figure 7.22: Compute Magnate Simulations – The distribution of the contribution levels of all users at the end of a year (disinterest factors obtained from a Poisson distribution of λ = **20**). At this unrealistically high simulated resistance level, nearly a third of the user base contributes less than 20% of their potential with over 1200 users not contributing at all. Despite this we see there is still a number of users contributing sustained by the social pressure from compute magnate incentives.

The results from this set of simulations show that the Compute Magnate incentives encourage increased user contribution of computational time.

Chapter 8

Conclusions

The major purpose of the Social Cloud for eResearch is to increase the uptake of public eResearch, or volunteer computing, through social influence applied via social networks.

This involves identifying three different roles that incentivize users by rewarding contributors. In this thesis, we have presented three roles:

- The Social Anchor role is awarded to those who are most active at bringing new recruits into the social cloud.
- The Compute Magnate role is awarded to those who both through their own contribution, and that of their friends, bring the largest pool of resources into the entire social cloud.
- The third role is the Project Champion, which is awarded to those who contribute strongly to a specific project.

In addition to these roles, we have introduced interest signatures, making it easier for users to choose projects based on the interests of their friends. This will also contribute to the formation of communities embedded within the Social Cloud.

We have presented an architecture that acts as a Project Manager and integrates Facebook, BOINC project servers and clients. This exploits an existing mature middleware platform (BOINC), and Facebook – the world's largest social network.

We envision that with the unique strengths of this approach to volunteer computing we will see significant increases in the processing power available to BOINC based projects. If we are able to recruit 1% of the current user base of Facebook to become BOINC volunteers, we will effectively double or even triple the present number of BOINC volunteers. Given some of the vital areas of research that many of the BOINC projects are involved in, we think this is a goal worth achieving.

8.1 Contributions

I architected and developed the Social Cloud for Public eResearch to combine a social networking service and a volunteer computing middleware, specifically Facebook and BOINC. In this thesis, I have described how the various components including Facebook, the BOINC project servers, BOINC clients and the core Social Cloud application interact.

I introduced a number of social engineering concepts and algorithms to meet the goals of the Social Cloud:

- Interest Signature – that helps quantify the interest areas of a user,
- Project Signature – that helps quantify the properties of a project so that users can discover projects that are most suited to their interests,
- Project Champion – that ensures smaller projects get more computational time and support,
- Social Anchor – that helps grow the number of volunteers,
- Compute Magnate – that helps ensure users are socially influenced to contribute with regularity,

- Social value – that helps identify which users are likely to help grow the Social Cloud faster, and,
- Compute value – that helps identify which users might not be achieving their potential to contribute.

All these concepts were proven to work in a series of experiments and simulations in chapter 7.

Through these novel and proven concepts, volunteer computing should see a significant growth in adoption in the future, and meet the previously mentioned goals:

1. Ease the process of an interested party becoming a volunteer.
2. Enhance the visibility of lesser known BOINC projects.
3. Incentivize user involvement, contributions and platform growth.
4. Maximize the computational power available to researchers.
5. Bring science closer to the general public and make it more meaningful.

Bibliography

[1] J. Wilkening, A. Wilke, N. Desai, and F. Meyer, "Using clouds for metagenomics: A case study," in *CLUSTER*, pp. 1–6, 2009.

[2] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good, "The cost of doing science on the cloud: the Montage example," in *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, SC '08, (Piscataway, NJ, USA), pp. 50:1–50:12, IEEE Press, 2008.

[3] Anderson, D.P. and Fedak, G., "The Computational and Storage Potential of Volunteer Computing," in *Cluster Computing and the Grid. CCGRID 06*, vol. 1, pp. 73–80, May 2006.

[4] BOINCstats, "BOINCstats | BOINC combined - Credits Overview." `http://boincstats.com/stats/project_graph.php?pr=bo`, June 2011.

[5] Facebook, "Statistics | Facebook." `http://www.facebook.com/press/info.php?statistics`, June 2011.

[6] Kyle Chard and Kris Bubendorfer and Simon Caton and Omer Rana, "Social Cloud Computing: A Vision for Socially Motivated Resource Sharing," *IEEE Transactions on Services Computing*, vol. 4, no. 4, 2011.

[7] BOINC, "Project List - BOINC." `http://boinc.berkeley.edu/wiki/Project_list`, August 2011.

[8] K. John, K. Bubendorfer, and K. Chard, "A Social Cloud for Public eResearch," in *proceedings of the 7th IEEE International Conference on e-Science* , (Stockholm, Sweden), December 2011.

[9] R. Perrott, "Infrastructure, requirements and applications for e-Science," in *Computer Architecture and High Performance Computing, 2002. Proceedings. 14th Symposium on*, pp. 3–10, 2002.

[10] A. Reinefeld and V. Lindenstruth, "How to build a high-performance compute cluster for the Grid," in *Parallel Processing Workshops, 2001. International Conference on*, pp. 221–227, 2001.

[11] Top 500 Supercomputer Sites, "November 2010 | Top 500 Supercomputing Sites." `http://www.top500.org/lists/2010/11`, June 2011.

[12] CCGrid, "Types of Cluster Computing | Cluster Computing." `http://www.ccgrid.org/types-of-cluster-computing.html`, February 2011.

[13] BOINC, "Why Use BOINC - BOINC." `http://boinc.berkeley.edu/trac/wiki/WhyUseBoinc`, August 2011.

[14] C. Catlett, "The Philosophy of TeraGrid: Building an Open, Extensible, Distributed TeraScale Facility," in *Cluster Computing and the Grid, 2002. 2nd IEEE/ACM International Symposium on*, p. 8, May 2002.

[15] Open Science Grid, "Learn About OSG." `http://www.opensciencegrid.org/About/Learn_About_Us`, August 2011.

[16] Amazon, "Amazon EC2 Pricing." `http://aws.amazon.com/ec2/pricing/`, August 2011.

[17] Eric Knorr, Galen Gruman, "What Cloud Computing Really Means." `http://www.infoworld.com/d/cloud-computing/what-cloud-computing-really-means-031`, August 2011.

[18] Microsoft, "Windows Azure Platform." `http://www.microsoft.com/windowsazure/pricing/`, August 2011.

[19] D. Anderson, E. Korpela, and R. Walton, "High-performance task distribution for volunteer computing," in *e-Science and Grid Computing, 2005. First International Conference on*, pp. 8 pp.–203, July 2005.

[20] Mersenne.org, "How it Works - GIMPS." `http://www.mersenne.org/various/works.php`, November 2008.

[21] E. Korpela, D. Werthimer, D. Anderson, J. Cobb, and M. Leboisky, "SETI@home-massively distributed computing for SETI," *Computing in Science Engineering*, vol. 3, pp. 78–83, Jan/Feb 2001.

[22] A. Beberg, D. Ensign, G. Jayachandran, S. Khaliq, and V. Pande, "Folding@home: Lessons from eight years of volunteer distributed computing," in *Parallel Distributed Processing. IPDPS 2009.*, pp. 1–8, May 2009.

[23] D. Anderson, "BOINC: a system for public-resource computing and storage," in *Grid Computing. Fifth IEEE/ACM International Workshop on*, pp. 4–10, Nov 2004.

[24] Kramer, D. and Maclnnis, M., "Utilization of a local grid of Mac OS X-based computers using Xgrid," in *High performance Distributed Computing.*, pp. 264–265, June 2004.

[25] Venkat, J., "Grid computing in the enterprise with the UD MetaProcessor," in *Peer-to-Peer Computing*, p. 4, 2002.

[26] F. Costa, L. Silva, and M. Dahlin, "Volunteer Cloud Computing: MapReduce over the Internet," in *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW).*

[27] Unofficial BOINC Wiki, "BOINC System Architecture - Unofficial BOINC Wiki." `http://www.boinc-wiki.info/BOINC_System_Architecture`, July 2011.

[28] Unofficial BOINC Wiki, "Resource Share - Unofficial BOINC Wiki." `http://www.boinc-wiki.info/Resource_Share`, June 2011.

[29] BOINCstats, "BOINCstats | BOINC Combined - User stats." `http://boincstats.com/stats/boinc_user_stats.php?pr=bo&st=0`, June 2011.

[30] BOINC, "AccountManagement - BOINC." `http://boinc.berkeley.edu/trac/wiki/AccountManagement`, June 2011.

[31] BOINC, "WebRpc - BOINC." `http://boinc.berkeley.edu/trac/wiki/WebRpc`, June 2011.

[32] Andrew Lipsman, "The Network Effect: Facebook, Linkedin, Twitter & Tumblr Reach New Heights in May." `http://blog.comscore.com/2011/06/facebook_linkedin_twitter_tumblr.html`, June 2011.

[33] Los Angeles Times, "Facebook F8: Redesigning and Hitting 800 million users - LA Times." `http://latimesblogs.latimes.com/technology/2011/09/facebook-f8-media-features.html`, September 2011.

[34] Neilsen, "August 2011 - Top US Web Brands." `http://blog.nielsen.com/nielsenwire/online_mobile/august-2011-top-us-web-brands/`, August 2011.

[35] Mark Gongioff, "Facebook Sucks Up a Ridiculously Huge and Growing Share of Our Time Wasted Online." `http://blogs.wsj.com/marketbeat/2011/09/26/facebook-sucks-up-a-ridiculously-huge-and-growing-share-of-our-` September 2011.

[36] Google, "Top 1000 sites - DoubleClick Ad Planner." `http://www.google.com/adplanner/static/top1000/index.html`, July 2011.

[37] Facebook, "Authentication – Facebook Developers." `http://developers.facebook.com/docs/authentication/`, June 2011.

[38] Facebook, "Permissions – Facebook Developers." `https://developers.facebook.com/docs/reference/api/permissions/`, September 2011.

[39] Facebook, "Graph API – Facebook Developers." `http://developers.facebook.com/docs/reference/api/`, June 2011.

[40] Facebook, "Open Graph Protocol – Facebook Developers." `https://developers.facebook.com/docs/opengraph/`, September 2011.

[41] Facebook, "Open Graph Beta – Facebook Developers." `https://developers.facebook.com/docs/beta/`, September 2011.

[42] Facebook, "Social Channels – Facebook Developers." `https://developers.facebook.com/docs/channels/`, September 2011.

[43] BOINCstats, "BOINCstats - BOINC Statistics / BAM! - BOINC Account Manager." `http://boincstats.com/bam/`, June 2011.

[44] BOINCstats, "BOINCstats - BOINC Statistics / BAM! - BOINC Account Manager - Detailed user, host, team and country statistics with charts for BOINC.." `http://boincstats.com/`, June 2011.

[45] GridRepublic, "GridRepublic | GridRepublic | BOINC Volunteer Distributed Grid Computing (About)." `http://www.gridrepublic.org/index.php?page=about`, June 2011.

[46] GridRepublic, "GridRepublic | GridRepublic | BOINC Volunteer Distributed Grid Computing (Projects)." `http://www.gridrepublic.org/index.php?page=projects`, June 2011.

[47] GridRepublic, "GridRepublic | GridRepublic | BOINC Volunteer Distributed Grid Computing (Community)." `http://www.gridrepublic.org/index.php?page=community`, June 2011.

[48] Intel, "Progress Thru Processors | Facebook." `http://www.facebook.com/progressthruprocessors`, June 2011.

[49] Foster, Ian and Kesselman, Carl and Tuecke, Steven, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *International Journal of High Performance Computing Applications*, vol. 15, pp. 200–222, August 2001.

[50] J. Howe, "The Rise of Crowdsourcing." Wired Magazine http://www.wired.com/wired/archive/14.06/crowds.html, June 2006.

[51] BOINC, "Choosing and joining projects." `http://boinc.berkeley.edu/wiki/Choosing_and_joining_projects`, August 2011.

[52] U.S. Securities and Exchange Commission, "Electronic Arts Form 10-K." `http://www.sec.gov/Archives/edgar/data/712515/000095013010001579/d10k.htm`, July 2011.

[53] U.S. Securities and Exchange Commission, "Zynga Form S-1." `http://www.sec.gov/Archives/edgar/data/1439404/000119312511180285/ds1.htm`, July 2011.

[54] Nick Wingfield and Anupreeta Das and Gina Chon, "Tech IPOs Test Sky-High Values." `http://online.wsj.com/article/SB10001424052702304447804576414111297459234.html`, June 2011.

[55] Unofficial BOINC Wiki, "How to decide on Resource Share - Unofficial BOINC Wiki." `http://www.boinc-wiki.info/How_to_decide_on_Resource_Share`, June 2011.

[56] K. A. M. and H. Michael, "Two hearts in three-quarter time: How to waltz the social media/viral marketing dance," *Business Horizons*, vol. 54, no. 3, pp. 253–263, 2011.

[57] Badgeville, "Badgeville." http://www.badgeville.com/about, Aug 2011.

[58] D. Takahashi, "By 2015, 50 percent of companies will embrace gamification, Gartner says." http://venturebeat.com/2011/04/14/by-2015-50-percent-of-companies-will-embrace-gamification-gartner-says/, August 2011.

[59] BOINCstats, "BOINCstats | BOINC Combined - Team stats." `http://boincstats.com/stats/boinc_team_stats.php?pr=bo&st=0`, June 2011.

[60] D. C. McClelland, *Human Motivation*. Cambridge University Press, January 1988.

[61] BOINC, "Computation Credit - BOINC." `http://boinc.berkeley.edu/wiki/Computation_credit`, July 2011.

[62] Unofficial BOINC Wiki, "Recent Average Credit - Unofficial BOINC Wiki." http://www.boinc-wiki.info/Recent_Average_Credit, July 2011.

[63] HCIL, University of Maryland, "Visual Analytics Benchmark Repository." http://hcil.cs.umd.edu/localphp/hcil/vast/archive/task.php?ts_id=119, 2009.

[64] M. Buchanan, *Nexus: Small Worlds and the Groundbreaking Theory of Networks*. W. W. Norton & Company, May 2002.

[65] The Economist, "Primates on Facebook." http://www.economist.com/node/13176775, February 2009.