A thesis
Presented to
The Academic Faculty
School of Sciences
Department of Computer Science and Engineering

By

Sylvester A. Terdoo

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science

European University Cyprus

January 2018

USING SOCIAL CENTRALITY TO ENHANCE DATA/STORAGE DISTRIBUTION IN AN EXTENDED ICN CONTENT DISTRIBUTION NETWORK

Approved by:

Dr. Katerina, Papanikolaou
Department of Computer Science and Engineering
European University Cyprus

Dr. Vicky, Papadopoulou
Department of Computer Science and Engineering
European University Cyprus

Date Approved: February 2, 2018

Learning never exhausts the mind

Leonardo da Vinci

A great dedication goes here.

ACKNOWLEDGEMENTS

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

A sizable percentage of the population today use online social network application to exchange different form of information and content over the internet. The recent growth of this online social networks has lead to the exploration of new paradigm that explore the content users desire to procure instead of the servers that make this content available, this paradigm is named Information centric networking, this paradigm focuses on a content-based approach instead of the traditional host-based approach. This new paradigm considers online social network behavior i.e. followers of user in popular media subscribe to content posted by the individual they follow, like how people subscribed to YouTube channel get notification of additional content once it is posted in the channel, or like how friends receive content from friend in Facebook. Here we recommend the use of Social Network Centrality alongside Content centric networking to help improve user access to content as well as maximize the use of their network resources. We show how Online Social network analyses can be used with Content centric solution, and the simulation we perform highlights the benefit of using this approach, how this approach improves upon the regular Content Deliver Network IP-based classification used in the distribution of content. We use Facebook data, mapping users to a network graph, using the nodes social proximity or centrality to determine where content should be cached, users of the network having access to this data using the relative shortest path will improve the overall network performance and improve the time needed to access content.

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Overview

Social network analysis SNA is not a formal theory in sociology but rather a strategy for investigating social structures. which uses mathematics to map and quantify the interconnection between actors in a network, an idea which is studied and applied in many fields. In fields like computer networks the aim is to understand the structural relationship between nodes as well as explain why this relationship occur. Moreover, the internet represent a social network of unrivaled scale. This interdisciplinary academic field has many advantages, as it equips researcher and scholars with tools with which to prove theory and study social network structures.

A principal guiding assumption in Social Network Analysis is that the medium used by a group to communicated to each other affects some important characteristics of the group. this can be measured on how effective the group are in performing particular tasks[1]. This is in contrast with the part of social sciences that assumes actors acts and make decision with no regard to the behavior of other actors.

Studies on off-line social networks have yielded fascinating properties such as the six degrees of separation effect or 'small word'[2], high clustering coefficient[3] or scale-free effec[4]. Others have also showed similar phenomenon in their online social network studies, realizing pattern that are explained by the scale-free effect and small world most notably Biehler and Leskovee[5]. The important element considered from Milgram research 1967 is that the majority of members in a social network can be assumed as a single graph.

In graph theory and network analysis, the concept of centrality was developed by people who sought insight as to how social power is acquired within a group. which they emphasized by saying that an individual does not have power in the abstract, they have power because they can dominate others. Centrality measure help identifies the important node within a graph or network. This process of identifying important node within a graph have various application which included identifying import infrastructure nodes on the network, network like the internet, it also is used to help identify key spreaders of content as well as studying how information go viral, it also be used to discover various efficient transmission route in a network.

Centrality measures can be used to measure popularity in a network, Nodes identified to be popular have high access to others and have a large number of other nodes willing to share information with them[6]. However high popularity and high influence don't mean the same thing which means that influence does not mean popularity and popularity does not imply influence.

Four basic measures of centrality have been developed over the years, from which every other sub variant has been derived from, each one exploring the various characteristics that makes a node central in a network, this four measures are:

1.2 Aims and Objectives

With various people using Online social networks (OSN) to communicate all over the world, and with recent growth of the Information centric networking (ISN) paradigm which involves the identification of data based on it name instead of it physical location and relies on in-network caching to store copies of data in the network, we adopt the Content Centric Networking(CCN) architecture of ICN and propose using social information in the design of a new caching strategy for a content centric networking that extends storage capabilities of CDN. We simulate how using centrality can be used to analyze a social network to determine possible location where content can be cached in the network. The main research contribution will include, how content storage can effectively be distributed over a CCN Network that extends the storage capability of a CDN.

1.3 Structure of thesis

This paper is organized as follows, first we outline the literature of Online social network analysis(OSN), the literature of the different centrality measures used to analyses online social networks, exploring their various applications, next we introduce the problem statement and the explain the current limitation of CDN networks and explain how information centric networking and social network analysis can be used to improve the performance of this well adopted networks, following this we introduce the proposed framework we have adopted to go about doing this, then we introduce how social network can be modelled as graphs and we explain how using the ICN paradigm can improve the performance of content storage in a network. Next, we describe how we simulate using Facebook data the principles of content-centric networking and execution of ICN. Finally, we conclude the paper and specify our future work.

1.4 Summary

Online social network like Facebook, Twitter, YouTube is all the rage this day, this network connects a great deal of people together. The result of the connection and communication of people on this social network continue to influence people personal lives as well as their business. Completely changing how people get and consume data, the availability of popular content on networking platform, people are effortlessly updated on current events. This network of how people communicated has therefore influences how content on the internet tend to be distributed. Popularly still adopted are Content delivery networks(CDNs), which provide user with better experience when accessing contents. This service uses host to host IP principles to distribute content in the network and fails to consider social semantics when distribution or transferring content to improve content caching. Information centric networking(ICN) is a popular new paradigm that has continued to be researched by the community, exploring how content that user require can be stored and delivered to them, replacing the host-host communication required to locate and deliver content user want. Popular content exchanged in social network includes photos or videos posted by individuals. Here we simulate how we can leverage information from this user network to improve storage performance of the network, reducing the latency delay and bandwidth requirement needed.

CHAPTER 2
FUNDAMENTALS

2.1 Introduction

In society today we see various different networks all around us, from rail line that connects cities to the cables that connect networks together forming the internet, from airline routes that connects countries and cities to brain neural pathways, ”Networks” is a simple term mostly used to to describe such structures. With millions of people all around our communities learning what this term means as well as using it in their everyday conversation, the complicated task of analyzing this network has just recently began to possible with the advancement in today computers, today in various different fields, various research use this term to refer to various different abstract structure with paths and point of intersection and are able to model various relationships edges between nodes, nodes referring to the cities, computers while the edges referring to the rails and cables. Various area of sciences that use mathematics like Graph theory continue till this day to analyses and study the various features of this structures and their various applications in our societies, this research has lead to the creation of sub fields like complex network analysis. Today in various different fields networks of various different kinds are represented as graphs.

2.2 Literature Reviews

2.2.1 History of social Network?

As early as the 80’s people argued among themselves that various social groups could exists having direct social ties that linked various individuals with similar belief and values. Georg Simmel a German Scientist in the 1980s was among the first to write directly in terms of social network terms, he introduced the fundamental concept of social networks in his essay which pointed out the nature of networks, at the time no advancement occurred, but as time past other researcher began investigating the concept, and how the idea of social network could applied in various fields like psychology, sociology, economics, mathematics and statistical physics[7], other scholars like Moreno began investigating the systematic analysis of social interaction in small groups, and W.Lloyd Warner explored various interpersonal relationships that happen in a network. today the concept of social network is used in a wide range of field having various meaning and is used to mean different things in different context.

As of 1970, many scholar began working together to combine the various knowledge of social network accumulated at that time, among one of this group had people like Charles Tilly, who focused on networks in political and community sociology and social movement, as well as Stangley Milgram, who later developed the ”six degrees of separation” thesis. Other research group most notably the group centered around Linton Freeman began investigated the topic social networks and it various mathematical application

Here we use Social network to refer to a sequence of node interconnected by one or more relation[8] that showcases the relationships between this nodes, how they communicated and support the overall network structure. Depending on the literature, any of the terms like vertices, players, node, links, edges will be used to represent actors and their relationship in a network, also depending of the context this network could represent terrorist activities around the city, paper published by a particular journal in the city

Borgatti, Mehra, Brass and Labianca [9] found in 2009 that within a social network they where three type of relationships between the actors in the network: the first kind of relationship is based on the interaction involves the constant interchange of messages between two parties or nodes like communication two friends communicating with one another through chat or speaking to one another, the second kind of relationship involves two nodes having similar attribute like their location within the network, and the final kind of relationship that can happen between nodes is one based on their social relationships, it is usually characterized by one node ability to learn about it from it environment , example will be like a node monitoring its resources like it memory.

2.2.2 Social Network Modeling

Social network are modeled using Graph, the first every use of graphs was by mathematician Leonard Euler in 1736, where he used graph to solve the Konigsberg Bridge problem , the question the problem pose was whether it was possible to walk through the city of Konigsberg[10] once and only once, which occupied two islands,was divided into four landmasses by its rivers and was linked by seven bridges. Euler reduced this problem to node representing landmasses and edges representing the bridges, and was able to prove the problem had no solution. While Euler initial study and use of graph mapped a relatively small abstract structure, modern day problem involves mapping complex structure like the relationships between various different network that make up the internet. While the scale of today problems differ in scale, the tools and thought process involved in solving them fundamentally remain the same.

Today in the field of Computer Science a graph G is defined as a triple consisting of a vertex set $V(G)$, an edge set $E(G)$, and a relation that associates with each edge two vertices(not necessarily distinct) called its endpoints, represented as $G = (V, E)$ [10]

A graph G is considered to be undirected when all the edges are bidirectional and the vertices of the graph are unordered, each vertex v in such a graph can be qualified by it degree denoted $deg(v)$, is value value is the number of edges incident to the vertex, with loops on such a vertex counted twice. A walk is a sequence $v_0, e_1, v_1, ..., v_k$ of graph vertices v_i and graph edges e_i such that for $1 \leq i \geq k$, the edge e_i has endpoints v_{i-1} and v_i[10]. A path is a trail which all vertices included in the walk is distinct, expect possible the first or last. The distance between two vertices in a graph is the number of edges in a shortest path connecting them, this distance is sometimes called geodesic distance. There may exist more than one shortest path between the nodes. A graph G is considered to be connected when there is a path between every pair of vertices, in such a graph every vertex in the graph is reachable. A cycle is a graph that consists of a number of vertices connected in a closed chain, with at least distinct edges in which the source and the target are the same. The cycle graph with n vertices is called C_n

A digraph$G$ is a directed graph whose edges are formed in ordered pairs of vertices such that $(u, v) \in E \neq (v, u) \in E$. In a digraphs every vertex in the graph can have either an in-degree or an out-degree. The in-degree α_D^{in} of a vertex v is giving by the number of head ends adjacent to v., while the out-degree α_D^{out} of a vertex v is the number of tails ends adjacent to that v.aka the branching factor . the various ideas and concept like cycle, path are extended to directed graph. while the distance between two node can also be measured in a digraph its not the same as the distance between two nodes in an undirected graph.

The neighborhood $N(v)$ of a vertex v in a graph G is the induced sub-graph of G consisting of all vertices adjacent to v but not include v itself.

A graph G can be implemented using a two-dimensional matrix, in this matrix each of the rows and columns represent a vertex in the graph. the value that is stored in the cell at the intersection of row$v$ and column w indicates if there is an edge from vertex v to vertex w. we can call this vertices adjacent vertices. Adjacency matrix $A = (a_{uv}$, is the starting point of the centrality computation, it properties are essential and will be discussed further in the section about centrality, this

2.2.3 Network properties

Today network a made up of a number of nodes which determine its overall size, a graph is considered to be a dense graph if the number of edges that comprise the graph is close tot he maximal number of edges possible in such a graph. while in contrast a graph with only a few edges in considered to be a spares graph. The distinction between the two graph are usually ambiguous and as such this graphs can appear differently depending on the context. a different quantitative measure used to measure and study complex network is called reciprocity. Formally reciprocity is defined as the probability vertices in a directed network are mutually linked. The clustering coefficient of a specific node in a graph is a measure of the degree to which nodes in the graph tend to cluster together. this basically is the likelihood that two randomly selected friends in a network are also friends in an un-directed graph representation of the network. The Homophily of a graph showcase the concept of ”Birds of a feather flock together” in a network, this basically say that people who share common characteristics like race, language and ethnicity are more inclined to relate and communicate better together.

2.3 Centrality in Social Networks

Measuring individual influence of actors in a network is a conceptual issue that has motivated the creation of various different measurement criteria used to measure a node influence in a network, various of this method differ from one another and use different yardstick for measuring influence, while many different people have different view of how centrality in a network can be measured, they all agree on the impact an influential node can have on a network. And thus an influential node in a network is categorized into two type; this is usually depending on the context. This actor can either be considered to impact the spread of information: people who influence other people , or this actor can showcase some combination of various desirable attributes like trustworthiness, expertise or network attributes like connectivity.[11] this actor referred to as an influencers have different names in different literature, some refer to them as key-players, other call them prestigious or spreaders etc. This day various different business often find it beneficial to find, access and quantify the prestige of spreaders in their network, since various of this actors are able to reach large number of audience with a very little effort or marketing cost, particularly using today’s technology. The process of discovering the various influential actors in a network requires the use of the various developed centrality measures(CM).

2.3.1 What is Centrality?

One can only determine what centrality does[12]. the different measures of centrality can be used to discover important node as well as quantify the extent to which this central nodes interacts with the other nodes inside the network. The higher the number of nodes a particular node is connected to the higher it's centrality[13]. Different researchers have been able to use centrality measure as an indicator to identify important actors in a social network and explain how this actors abilities are connected with the network[14]. Many proving the important of the measure of centrality in social networks.

Wasserman & Faust demonstrated that central nodes in a network are more active in order to manage their own contacts, they do this by trying to reduce the paths with other nodes in the network, which means increasing the number of direct links they have with other nodes. The node in this position have easy access to information and resources and hence are able to better control information spread in the network. Another method developed which attempts to quantify and identify important vertex in a network is Social capital, but in contrast to centrality which tries to identify the important nodes in the network, this method focuses on the features of the network that contribute to the individual node[15].

2.3.2 Measures of Centrality

As highlighted previously various different measures of centrality have been developed over the years, However they are four basic highly developed measures, that other measured are derived from, this four basic centrality measure include: Degree centrality , Betweeness centrality, Closeness centrality and Eigenvector centrality which employs the similar concept used in the Google pageRank algorithm. The basic definitions of this measures and their various application in the various different type of graph, as well as the formulation and interpretation are discussed below.

Degree Centrality

This measure was introduced in the field of graph theory, degree centrality (DC) σ_D captures the a node contentedness in its neighborhood(direct hop distance). A vertex i is considered a hub in a network if its centrality score is high, meaning it has a large number of contact or edges [12]. In directed networks nodes normally have two separate measure of degree centrality, in-degree and out-degree. The degree centrality of a vertex v, for a given graph $G = (V, E)$ with $|V|$vertices and $|E|$ edges, is defined as

$$C_D(v) = deg(v)$$

the equation above however is impacted by the size of the network, if the result is to be compared with other graph, the formula it normalized by dividing by $n - 1$. this gives a probabilistic value between 1 and 0.

Closeness Centrality

This measure came about in the 1950s, it explored a different view in measuring the centrality of a node in a graph. it relates how one node in a network might control the various communication that occur inside the network, by keeping track of relative distances of node i to other nodes in the network. this measure is known to scale directly with distance i.e nodes twice as far, is half as central. Formally the Closeness centrality of a node can be defined as the average length of the shortest path between the node and all other nodes in the graph. in simple terms the central node in the graph have shorter access to the other nodes in the graph. Mathematically it defined as the reciprocal of farness[16]

$$C(x) = \frac{1}{\sum_y d(y, x)}$$

where $d(y, x)$ represent the distance across x & y. this formula normalized version is gotten by multiplying the above formula by $N-1$, where N is used to represent the number of nodes in the graph. this normalization allow graph of various different sizes to be compared

Betweenness Centrality

First introduced by Linton Freeman, he described it as a measure for quantifying the control of human on the communication between other humans in a social network. to quote him directly 'a point that falls on the communication paths between other points exhibits a potential for control of thier communication paths between other points exhibits a potential for control that defines the centrality of these points'[17]. this measure of centrality is used to ascertain the size of crossing point in between routes, a designated node is central if it lies in maximum number of shortest paths connecting distinct pairs of nodes on a network. Formally it is defined as quantifying number of times a node acts as a bridge along the shortest path between two other nodes. Mathematically betweeness of a nodes is expressed as the ratio of $\frac{g_{jk}}{g_{jk}}$ for all pairs of node, formally

$$C_B(i) = \sum_{jk \neq k} \frac{g_{jk}}{g_{jk}}$$

where $g_{jk}(i)$ represent the number of shortest paths connecting jk passing through i , and g_{jk} represent the total number of shortest paths

The above formula is normalized by diving by $\frac{(n-)(n-2)}{2}$, which represents the number of pairs of vertices excluding the vertex itself.

Eigenvector Centrality

Introduced by Bonacich, it is also know as EigenCentrality. This measure showcase the idea that a node is only important if its neighborhood is also important i.e it's adjacent node centrality value factors into determining how important a node is. Unlike degree centrality which depends on having many connections regardless of whether this connection lead to isolated nodes, EigenCentrality measures says a central node should be connected to other powerful nodes. a well know variant of this centrality measure is the Google's PageRank algorithm. Mathematically eigenvector centrality can be described using adjacency matrix as follows:

for a given graph $G = (V, E)$ with $|V|$ number of vertices,
Let $A = (a_v, t)$ be the adjacency matrix.. that is (a_v, t) has the value one(1) if there is a link between v and t, it value is zero(1) when v and t are not connected. the equation below defines the centrality score of v

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

with the first part of the equation representing the Neighborhood of x_v , the second half representing the adjacency matrix of the graph and λ is a constant. the equation can be written in vector notation as

$$Ax = \lambda x$$

2.4 Application of Centrality

1. Centrality measures are used for understanding Information spread / diffusion
   - It is used to determine the extent information has spread ?
   - Who are the most influential in spreading information ?
   - Understand how does a top/ video become viral?
   - Reveal speader of infectious content in socially interactive network
2. Centrality is used in networks to Identify influential users / experts
   - Recognize mportant nodes acting as focal point in technological network.
   - Identify most influential personality or event in social network.

- Identifying topical experts
- Detect key strategies in a business-oriented network
- find purveyor(a person or group who spreads or promotes an idea) of content in a chain network

3. Centrality is used by businesses and Social media for Search and Recommendation
    - Social media like Twitter use the measure to find people with similar tastes and recommend this people as friends
    - Business like Amazon use this measure to recommend books to people based on other similar customer tastes
4. Centrality is used for opinion mining
    - Identify people opinion on various topic
    - Used to summarized opinions like based on people opinions online how is most likely to win an election
5. Centrality is used for Spam detection
    - Used to identify users with malicious intentions as well as identify spams in social networks.
    - used to identify trustworthy entities, through mechanism like ratings.
6. Mining information on recent events
    - Used as valuable sources of information on events that are happening currently. example it has been found that events like earthquakes are know to people in a relative area through tweets before they feel the tremor. as people instinctively tweet about such events.
7. Several Centrality-based classification can be used to:

- Discover routes for efficient transmission.
- Measure cohesiveness between nodes in a network.

2.5 Information centric networking

The fundamental concept of Information centric networking according to [18] is a set of interconnected pieces of information which can be addressed by their names for routing and managed by applications for services. This could be static or dynamically generated, real time video stream etc. The primary concern of the network here is to distribute , find and deliver information. This approach is drastically different from IP which established communication between two hosts before content is transferred, the delivery of data in this network follows a source-driven approach. This approach uses a receiver-driven principle which means the users request content without knowing the host who will deliver the content, the network is responsible for mapping the request to where the requested content can be found, this idea of matching request to content rather than finding which endpoint can provide the content is key and this separation of naming from location is the fundamental idea of ICN. To be able to do this separation, the naming of content is done in such a way that its independent of the location the content resides on. This feature and the fact that the network has a native caching function that gives nodes in the network the ability to cache content for a while and be able to deliver them to requesting users hence enabling efficient delivery of content to users. This idea of separating location from content name also support mobility. Even when users relocate and are connected to another node in the ICN network, since no IP address is used for routing nothing is affected, unlike the case of IP where the addresses needs to be updated. In an ICN network once content is requested it is cached and delivered to future request.

CHAPTER 3

RELATED WORKS

How content can be Cached, has been researched by different research communities, employing different techniques like small cell Networks(SCNS), this technique employs a network model where various small bass station(SBSs) are able to cache content, this process helps regulate the load on edge servers in a network, providing users with an enjoyable experience and satisfying their different demands overall. Small bass station is placed randomly using a process called Poisson point process. This distributed edge servers provide the content from the internet or by serving them using a locally cached copy. [19][20], many other professional user network properties like homophily which assumes birds of the same feather flock together to explore how content can be stored in an efficient manner in a complex network like a Content Distribution Networks(CDNs) , using game theory to formulate the caching problem as a many to many matching problem, modelling the small base station and service providers. Attacking the caching problem using a much different point of view. The decision of caching content in local base station is done to reduce the delay experienced by users, this base station use the concept of local popularity of the video content to determine which videos are to be saved, this helps reduce the overall load on the back links of the overall network. [21]. In [22], the decision of managing how to effectively cache replicated content across various dedicated storage devices attached to nodes using ICN. Also, in [23] the problem of caching under constrained storage capacity across routers distributed in the network. Researchers have found that network topology and content popularity are two crucial factors that must be considered to ensure content are effectively cached. [24] looked at the idea of pushing content to the edge to anticipate network congestion, while [25] basically computed the volume of an ad-hoc network can cache.

In more recent work [26], various game theory concept is utilized for caching popular videos at small cell base stations(SBSs). In other similar works [27] proposed an approach for ICN that simulate how wireless access point owners can jointly rent their unused storage space to content provider under partial coverage constraints. Both papers instead of providing an efficient content placement solution, mostly targeted a pricing model.

In his paper [28] defines a conditional betweenness centrality, basically using mathematics to map and quantify the interconnection between actors in a network, an idea which is also used here, this metric determine which node in the network will cache content. In [29] a new social-aware metric which is adaptable to dynamic network topology to cache content at vehicle. [30] uses social information in Content Centric Networks (CCNs), this Social-Aware Caching Strategy(SACS) pro-actively cache content produced by influential users in a network, using Page-rank and Eigenvector centrality measure to identify this influential/ prestige users.

Also in [31], centrality based caching approach is used in CCN, where the content to be stored is based upon its centrality. The authors exploit different centrality like graph eccentricity, degree, stress, closeness and betweenness centrality to miscellaneous decide edge server content should be saved. Its begin proposed that a simple degree centrality based allocation is sufficient to decide which edge server content should be stored also others have proposed that content with high cache hit rate are to be cached at edge servers with higher betweenness centrality. Here we use a metric that determine where content can be caches based on the concept of homophily which says, birds of the same feather flock together. in our case, groups with the same social interest are bound to be interested in the same content, using the social proximity of individuals within a group, we determine where in the network content could be cached, minimizing the overall distance between the cached content and the various individuals within the group.

3.1 Introducing the problem

Today the internet has affected the way people communicate and do business, for different business the effectiveness of this communication is vital for the business to remain afloat, till date various effort continue to be made to improve the overall time over this network. Naturally one would expect that increasing the bandwidth of internet infrastructure will increase the capacity of the server hence reducing the access delay to subscribers, however this is not the case, an increase in the bandwidth of internet infrastructure will not necessarily improve subscribers access time problems.

Two main occurrences result in bad services quality over the internet, one is the overall lack of management over the internet, the other is the fact that the amount of contents in use over the internet is unimaginable large. The two reasons not only makes access delay a problem over the internet, but makes the access latency unforeseeable. The approach adopted to address this problem is to have this content moved from their initial origin servers to edge servers closer to the users. Using edge servers to serve use content normally allow for higher content transmission rate and much lower access latency when compared to the when users a served with content from the original server, and the approach of using replicated edge servers to provide this content to users reduces the cost of access even more than having one of such servers as various users are directed to the various different replicated servers, enabling load balancing among this servers. CDN takes this approach by replicating a set of content on the replicated edge servers, sending request for replicated content to the appropriate replicated servers. However deciding how to place content on this replicated edge servers which distribute content to users remain a key challenge in designing an effective CDN, this along side a few others remain one of the key challenges faced by providers like Akamai how provide such servers. Instinctively, it would occur to have this edge servers placed close to the users, but how exactly could this be done

3.2 Overview

CDN are used to get content closer to users, this is done by distributing content to different replicated edge servers placed closer to users. Content provided by this edge servers are content requested by some user at a particular point in time, this requested content stored on the edge server a content with high hit ratio from users. The overall process ensures that users have access to content very quickly with as less delay as possible and utilize a small amount of bandwidth to access this content. This particular feature of CDN edge servers make them very coveted by popular content distributors on the internet, providing servers for multiple content distributors using shared resource make this network vital in today's internet.

3.2.1 Distribution System

CDN distributes the contents from the origin server to replicated edge servers which are much closer to end users. The implication of this is that provide to users little to no access delay and use overall require users to use less network bandwidth. The internet remain the most commonly used approach to distribute contents to various edge servers, a CDN using this approach maintains an overlay of its distribution tree over the existing Internet infrastructure and disseminates content from the origin server to the replicated edge servers via the tree. Establishing and maintaining this overlay or tree remain a major CDN technical concern. if a generalized approach is used in this overlay, performance issues may arise.

3.2.2 Replicated Edge Server

Deciding where on the internet to place content, so that clients content access latency as well as the network bandwidth use will be minimized remain the problem, incidentally this two goal have a mutual relationship, such that improving the use of one will lead to improving the use of the other.

Effective content placment involves using replicated edge servers all over the internet, various sound approaches have to theorized to model this problem. Many of which is based on the center placement problem. Because of the complexity of these algorithms, various different heuristics have been proposed which tries to improve upon this algorithms.

Theoretical Approaches

The content placement problem is usually modeled as a metric k-center or metric facility location problem. This is an NP-hard optimization problem studied in theoretical computer science, it is specified as follows: Given n cities with specified distances, one wants to build k warehouses in different cities and minimize the maximum distance of a city to a warehouse.[32] Another similar type of this problem in computer science is the facility location problem which is concerned with the optimal placement of facilities to minimize transportation costs. [33], problem is also considered NP-hard to solve optimally. Another graph theoretical approach is the k-hierarchically well-separated trees(k-HST)[34][32].

Heuristic Solutions

The above theoretical approaches are know to be either computationally expensive or do not take into account the overall features of the network or workload required. Making them unsuitable for practically used in CDN systems[35]. Some heuristic algorithms have been proposed, this algorithms are designed taking into consideration the network topology or similar observed paths from other existing networks[36][33][37][35], this approach provides solutions with much lower computational cost.

Greedy Algorithm

P. Krishman[36] for the cache location problem proposed a greedy algorithm which was later adapted by Qiu et al.[33] for the server placement problem in CDN.

This greedy algorithm after every iteration chooses one site from a set that satisfy certain constrain. Generally the algorithm assumes that client direct their access to the server closest to them, this measure to be like the server that can be reached with the lowest cost. The process of the algorithm goes as follow: Imagine out of S sites, M potential servers needs to be selected, the first iteration in the algorithm examines all of the S site, checking to see their suitability to host a server. the cost associated with each site is computed under the assumption that every client request end up at that site, it selects the site that yield the lowest cost, like the lowest bandwidth cost. This process of computing the cost continues in every iteration, each time the selected site yielded the lowest cost in conjunction with the previously selected site, this pattern continues until the desired number of M servers have been chosen.

Another algorithm similar to the one described, but various differences was all discussed, this algorithm uses l-backtracking, this algorithm is called l-backtracking greedy algorithm, this algorithm unlike the previously described algorithm checks after each iteration all the possible combinations achievable by removing l of the already placed servers and replacing them with $l + 1$ new servers. Going by this the previous algorithm is a 0-backtracking greedy algorithm.

The above greedy algorithm work relatively well, but have a major draw back which is the fact that the greedy placement algorithm requires information of the client location in the network and all pairwise inter-node distances, which in many is not available.

Topology-informed Placement Strategy

Transit Node as this is formally called is a kind of topology-informed placement approach discussed by [37]. its stated as follow, nodes selected as possible host are node with high outward degree, this is done under the assumption that this node have access more nodes with low cost. nodes are selected from highest to lowest number of outward degree. the drawback of this heuristic is the lack of more detailed network topology, it operates under the premise that nodes in the core of the internet transit points will have high degree, and node link corresponnds to AS-level BGP peering. A slightly better is suggested in[35], this approach utilizes the router-level Internet topology, rather than the AS-level topology. i.e this using local area networks associated with a router as a possible site to place this edge server, instead of using AS exchange point as possible site.

This above describe methods have been found to work equally as well as the greedy placement, and even better measurement conducted has shown that using router-level topology information placement result in a more better performance than all other approach discussed approaches like the different greedy algorithm an the AS-level topology information. But performance issues arise when servers are placed indiscriminately, i.e when high number of servers are placed in the network the performance of this network begin to shrink or decrease. because because this capacity of a single host site, have limited capacity, the number of them

Finding the sweet spot where content can be placed in the network remain till this day remains a problem and is considered to be a problem that is not well explored from the various literature available.

CHAPTER 4
TECHNICAL APPROACH AND EVALUATION

4.1 Proposed framework

The Content Centric Networking (CCN) paradigm is considered the future architecture for the current internet, few solutions have been proposed that leverage information about users social communities to design efficient content replication schemes. CCN approach uses in-networking caching to improve users experience when accessing data. Online Social Networks (OSN) have become an important field in understanding how communication occur in a network of user alas how information or data is spread between individuals in this network. Using the knowledge of Online Social Network Centrality and Information Centric Networking CCN paradigm, CDN Information distribution and retrival can be improved.

Here we show a caching strategy for a CCN extended CDN network which considers the social semantics of the network when distributing or transferring content to be cached. Using Online Social Network Centrality Measures we identify central users in a network - people that are central to the various activities happening in the network(i.e act as bridge, receive the most attention from other users and whose interest are more likely to be explored by the many that follow or depend of them). The strategy exploits the social proximity of this central users to other users in the network, using this relationship key strategic location where content can be cached will be identified and recommended.

Using simulation and data on user from online social network services like Facebook, we demonstrate the value of using centrality do determine where in the network certain information type can be cached, taking into consideration the centrality of users interested in the data and the proximity of other interest users to the central users.

4.2 Network model

This section will illustrate in detail the model used, we assume the CDN network architecture extends content centric networking. As the internet continue to evolve as more of a social oriented network. We propose a social network model built over thee CCN network, modeling how user interact over the network. Using a use case we model the interactions of users in a social network built over a content centric network. We point out the limitation of current CDN networks, showcasing how the idea of social centrality on an CCN extend CDN can improve retrieving and storage of data.

4.2.1 Content Centric Networking

CCN architecture is widely received among the various existing ICN architecture. This communication architecture uses two attributes namely Data and interest. Specifically user request content by broadcasting their interest message over the CCN network, among the nodes who hear receive this broadcast, any node with the requested content will respond with a data message. The caching of data messages by nodes is important as caching policies are used to regulate message caching as these caches have finite space. Ensuring that content is cached in strategic positions in the network so users can access them easily is the goal of this project.

4.2.2 Social Network Model

Online Social networks allow user to share different content with the world or their close acquaintances. Friends may also receive update about the activities of their other acquaintances, and may further choose to share this content continuing the circle, causing this content to spread around such groups. With this observation we model this relationship as a social network where users are able to either distribute of receive information from and to members of their communities, and to simulate the interaction of users in this social network, we employ social network analysis.

4.2.3 Socially-Aware Caching Strategy

Future internet will rely on CCN architecture, as more groups in social networks will continue to organize themselves and continue to exchange large amount of content through this network. We prose using the social information of users and centrality measures to perform content centric networking caching. This will take into consideration the proximity of users, the centrality of users relative to other users in the communities. This relationship will be exploited to recommend where in the network content can be cached, as an example a popular Youtuber with lots of followers and subscribers, when this person produces content online, many of the followers will most likely access this content and in turn might recommend this video to their subscribers or followers, this circle goes on.

In the strategy we propose using centrality and proximity relationships of users in the network to identify and recommend where data should be cached in the network, to allow user access to this data using the minimum distance. Hence improving the overall access of users to content in the network.

4.3 Influential Users Detection

Using Betweenness and Eigenvector centrality measures we detect users central to the network. This centrality measures like the others help measure the importance of users in a network. By using both of this measures we are able to identify users who can act as some sort of a bridge between content and the other users. We measure the overall average of every user in the network, those with the highest score are considered central. Hence the proximity of this central users to other users will be used in determine possible places content can be cached so every user can easily access this content.

4.4 Simulation

This section describes the simulation environment and the parameters used to adjust our model to evaluate our social aware strategy. Below we give a more in-depth and general description of our simulation environment.

4.4.1 Social Network Topology

To model the social relationship between users as a network we use publicly available dataset, the Stanford community provide a Facebook dataset, this dataset represent individual as node and an edge list as the relationship between this individual, this data comprises of 4, 039 users, 88,234 friends relationships and each user counts in average 44 relationships. Our simulation experiments is performed modelling this social network relationship as a graph. This dataset consists of circles (or friends lists) from Facebook. This anonymized dataset includes node features (profiles), circles, and ego networks. The edges are undirectected.

We consider a future content-centric internet where CDN networks extends the ICN CCN architecture. This network architecture is structurally equivalent to todays internet. The edges described in the data could be of any form: friendship, collaboration, following or mutual interest. Here we specifically study and build our model over Facebooks social network.

4.4.2 Simulation Tools

To evaluate our social aware caching strategy for this network, we implement a discrete-event simulation scenario in python Jupyter notebook. Using various open source libraries like, matplotlib a python plotting library for making publication quality figures in a different format across different platforms. plotly a graphing library for making interactive, publication-quality graphs. networkx a collection of network analysis tools with the emphasis on efficiency, portatbility and ease of use. Networkx is open source and free. numpy adds support for large, multi-dimensional arrays and matrices along with a large collection of high level mathematical functions to operate on these arrays and finally datascience an open source python library used for scientific computing and technical computing.

We demonstrate this caching strategy using online social network analysis measures like centrality which detect influential user in social network community, implementing a parallel betweenness centrality algorithm we demonstrate how this strategy can be used alongside already existing Content Centric Network caching strategies and replacement policies in an ICN network. Using this framework along side traditional replacement policies like Last Recently Used(LRU), Random(RAND) and FIFO etc.

4.5 Simulation Experiment

We exploit available social structures and employ already existing in literature graph representation of networks and proposes a method where data is stored on central nodes in the network because of their significance. This attribute stems from the extent to which they hold an intermediate position between the current node and the rest of the network nodes. The out of horizon demand is inferred by the value of the centrality of the node relative to other nodes in the network, introduced to quantify this significance and accordingly pick the nodes to take part in the local solution. Having exploited the centrality of nodes in a network, communication and computation cost are avoided as they is no need to apply any mapping mechanism for capturing this demand load. This solution is not guaranteed to always reach the optimal, but solution show that it achieves satisfying convergence, especially when applied on graphs incorporating with some real-world social characteristics. In the future we plan to extend this work to a more general solution, since the model studied here is constrained according to uniform demand hypothesis. Additionally, other attributes and characteristics such as correlated load demands which generally better model real-world social user nodes communication are characteristically opportunistically considered.

We propose a social aware caching strategy for content centric networks. This strategy uses information about users in a network by, determining where to cache content by considering the relationship between users. Based on extensive simulations experiments, we showed that our caching strategy improves the caching performances of CCN extended CDN over tradition CDN performances on the cache hit and stretch. Our objective is to evaluate the caching performances of the CCN extended CDN network architecture using the proposed social-caching strategy to the performance of traditional CDN architecture without the social caching strategy. Using the various social network analysis centrality measures specifically centrality we simulate the behavioral performance is different similar scenarios. For our purpose the result we like to show was that of be behavioral performances of the system when compared to the computed betweenness centrality measure of social network analysis. The other measure also show promising result in deferent situation.

4.6 Results

The result of our simulation experiment is presented in the figure below. Below are four figures. Two of each depict each of our metric: Cache Hit, Stretch. All the figures share the same axes: the x-axis is the cache size; the y-axis is the probability. For each simulation we performed 100 runs using different social network activity traces and provide the average value and the confidence intervals. Figure 2(a) and 2(b) illustrate the cache hit performance of the CDN network without the social aware caching strategy. Without the strategy, the cache hit of the CDN achieves low values and it barely reaches 5 percent. On the other hand, using the social aware caching strategy increases significantly the cache hit and reaches 30 percent using betweenness centrality measure with the Facebook data set.

The low performance of CDN are due to the use of large scale and realistic social realistic social and network topologies. Indeed, the long routes traversed by the content after the cache hit and its computation works as follows. In every hop passed by data or content, the cache hit gets updated with a hit or miss. If the requested content is present at hop n, we obtain a single hit and there have been n-1 hops without the requested content. As the cache Hit metric is the ration of hit about the number of miss, the longer the path to find the content is, the lower the cache hits it. In our case, our strategy pro-actively caches content on a central node on the network paths. This social aware caching strategy succeeds to make the content more available and improves drastically the caching performance of CCN extended CDN.

The Stretch metric is presented in figure ab. This metric is in direct correlation with the cache hit. Out using our strategy, content traverse the shortest path to get to the requester from a specified node. Using the social aware caching strategy, the distance to get the content has greatly redued, in this case content takes on average half the previous shortest path to reach a specific node. As an aside it also crucial to note that the number of user in the data set has not impact on the performance of this strategy.

4.7 Evaluation Results

Extending the CDN architecture with CCN-base architecture and social aware caching strategy for content storage and delivery, we can intuitively imagine that the network load can be largely reduced depending on the proximity of users to the central nodes in the network. Here we performed some evaluation to prove the advantages of a CCN extended CDN architecture. Compared to the existing current architecture, using many users on an online social network.

In our simulation, we took an example of friends relationship in a Facebook network. We consider the use of CDN with one server somewhere and one at the peering point of the social network. To compare with the CCN social caching solution, we assume both the same content is delivered from the CDN as well as CCN extended CDN network. For CDN to be effective the servers need to have a lot of contents; this prevents having CDN nodes close to end users. In the ICN CCN solution, the community and social proximity of user in their social communities are recognized and keep track of in their router table, whereas in the case of a CDN, it has to be provisioned using some defined choices, hence creating a trade-off between a poor hit ratio if the servers are close to end user and the need for huge databases if they are in the core network. For this study we used a two level CDN architecture with contents stored in a fixed location and a peering point for the second level. The latency to reach those nodes would be an intermediate value between what it would be to reach network devices among the different communities in the network. For the operator network topology, we assumed a hierarchical three-level network topology, with caching facility at each level of the CCN architecture.

CHAPTER 5
CONCLUSION

Resulting from the evaluation, using ICN CCN social aware caching for data storage and retrieval has many advantages for network operators as expected, it also provides superior quality experience to end user since the content is closer to these end-users and can be delivered more rapidly. Furthermore, since the social links derived from the Online Social Network have a direct influence on the network delivery efficiency and perfectly fit with ICN CCN features it will in the near future be heavily integrated to us to data internet network architecture. This ICN paradigm stand to become a promising solution for various networking applications. In our future work, we seek to better tune our described model as well as adapt an application that utilizes this caching solution as well as make large scale tests on a real network to compare with the current infrastructure. We also plan to explore social links to optimize the routing and delivery of contents over the internet.

Appendices

APPENDIX A

EXPERIMENTAL EQUIPMENT

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

APPENDIX B
DATA PROCESSING

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

REFERENCES

[1] J. Scott, *Social network analysis*. Sage, 2017.

[2] J. Travers and S. Milgram, "The small world problem," *Phychology Today*, vol. 1, pp. 61–67, 1967.

[3] M. E. Newman, "Analysis of weighted networks," *Physical review E*, vol. 70, no. 5, p. 056 131, 2004.

[4] M. Castells, "Informationalism, networks, and the network society: A theoretical blueprint," *The network society: A cross-cultural perspective*, pp. 3–45, 2004.

[5] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proceedings of the 17th international conference on World Wide Web*, ACM, 2008, pp. 915–924.

[6] R. Cross, S. P. Borgatti, and A. Parker, "Beyond answers: Dimensions of the advice network," *Social networks*, vol. 23, no. 3, pp. 215–235, 2001.

[7] J. Scott, "Social network analysis," *Sociology*, vol. 22, no. 1, pp. 109–127, 1988.

[8] ——, "Social network analysis: Developments, advances, and prospects," *Social network analysis and mining*, vol. 1, no. 1, pp. 21–26, 2011.

[9] S. P. Borgatti, C. Jones, and M. G. Everett, "Network measures of social capital," *Connections*, vol. 21, no. 2, pp. 27–36, 1998.

[10] D. West, *Introduction to Graph Theory*. Prentice Hall, 1996, ISBN: 9780132278287.

[11] E. Keller and J. Berry, *The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy*. Simon and Schuster, 2003.

[12] L. C. Freeman, "The development of social network analysis-with an emphasis on recent events," *The sage handbook of social network analysis*, vol. 21, no. 3, pp. 26–39, 2011.

[13] P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[14] R. S. Burt, "The contingent value of social capital," *Administrative science quarterly*, pp. 339–365, 1997.

[15] S. P. Borgatti, “Identifying sets of key players in a social network,” *Computational & Mathematical Organization Theory*, vol. 12, no. 1, pp. 21–34, 2006.

[16] A. Bavelas, “Communication patterns in task-oriented groups,” *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 725–730, 1950.

[17] L. C. Freeman, “A set of measures of centrality based on betweenness,” *Sociometry*, pp. 35–41, 1977.

[18] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, “A survey of information-centric networking (draft),” in *Dagstuhl Seminar Proceedings*, Schloss Dagstuhl-Leibniz-Zentrum fÃ1/4r Informatik, 2011.

[19] E. Baştu, M. Bennis, M. Kountouris, and M. Debbah, “Cache-enabled small cell networks: Modeling and tradeoffs,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 41, 2015.

[20] B. Bharath, K. Nagananda, and H. V. Poor, “A learning-based approach to caching in heterogenous small cell networks,” *IEEE Transactions on Communications*, vol. 64, no. 4, pp. 1674–1686, 2016.

[21] K. Hamidouche, W. Saad, and M. Debbah, “Many-to-many matching games for proactive social-caching in wireless small cell networks,” in *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014 12th International Symposium on*, IEEE, 2014, pp. 569–574.

[22] V. Sourlas, L. Gkatzikis, P. Flegkas, and L. Tassiulas, “Distributed cache management in information-centric networks,” *IEEE Transactions on Network and Service Management*, vol. 10, no. 3, pp. 286–299, 2013.

[23] Y. Wang, Z. Li, G. Tyson, S. Uhlig, and G. Xie, “Design and evaluation of the optimal cache allocation for content-centric networking,” *IEEE Transactions on Computers*, vol. 65, no. 1, pp. 95–107, 2016.

[24] Y.-T. Yu, F. Bronzino, R. Fan, C. Westphal, and M. Gerla, “Congestion-aware edge caching for adaptive video streaming in information-centric networks,” in *Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE*, IEEE, 2015, pp. 588–596.

[25] B. Azimdoost, C. Westphal, and H. R. Sadjadpour, “Fundamental limits on throughput capacity in information-centric networks,” *IEEE Transactions on Communications*, vol. 64, no. 12, pp. 5037–5049, 2016.

[26] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2115–2129, 2016.

[27] M. Mangili, F. Martignon, S. Paris, and A. Capone, "Bandwidth and cache leasing in wireless information-centric networks: A game-theoretic study," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 679–695, 2017.

[28] P. Pantazopoulos, I. Stavrakakis, A. Passarella, and M. Conti, "Efficient social-aware content placement in opportunistic networks," in *Wireless On-demand Network Systems and Services (WONS), 2010 Seventh International Conference on*, IEEE, 2010, pp. 17–24.

[29] J. A. Khan and Y. Ghamri-Doudane, "Saving: Socially aware vehicular information-centric networking," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 100–107, 2016.

[30] C. Bernardini, T. Silverston, and O. Festor, "Socially-aware caching strategy for content centric networking," in *Networking Conference, 2014 IFIP*, IEEE, 2014, pp. 1–9.

[31] D. Rossi and G. Rossini, "On sizing ccn content stores by exploiting topological information," in *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*, IEEE, 2012, pp. 280–285.

[32] S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang, "On the placement of internet instrumentation," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, IEEE, vol. 1, 2000, pp. 295–304.

[33] L. Qiu, V. N. Padmanabhan, and G. M. Voelker, "On the placement of web server replicas," in *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, IEEE, vol. 3, 2001, pp. 1587–1596.

[34] Y. Bartal, "Probabilistic approximation of metric spaces and its algorithmic applications," in *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, IEEE, 1996, pp. 184–193.

[35] P. Radoslavov, R. Govindan, and D. Estrin, "Topology-informed internet replica placement," *Computer Communications*, vol. 25, no. 4, pp. 384–392, 2002.

[36] P Krishnan, D. Raz, and Y. Shavitt, "The cache location problem," *IEEE/ACM Transactions on Networking (TON)*, vol. 8, no. 5, pp. 568–582, 2000.

[37] S. Jamin, C. Jin, A. R. Kurc, D. Raz, and Y. Shavitt, “Constrained mirror placement on the internet,” in *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, IEEE, vol. 1, 2001, pp. 31–40.

VITA

Vita may be provided by doctoral students only. The length of the vita is preferably one page. It may include the place of birth and should be written in third person. This vita is similar to the author biography found on book jackets.