Abstract

[Place holder abdtract]

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Using graphs as representation and algorithms that traverse graphs are common through many scientific domains[**?**]. We can use graphs to simulate how a disease would spread through a social network, we can predict how efficient viral marketing is and we can see how a new trend would spread through a community[**?**].To select the most influential nodes for a information diffusion through a social network is an NP-hard problem[**?**]. One of the common ways to select them, is by brute forcing through the network with algorithms like BFS. One of the solution is the greedy algorithm[**?**]. The greedy algorithm, proposed by[MaximizeSpread2003] goes through all the nodes and computes the effect on the network that nodes had. The greedy algorithm takes the k top most influential nodes as the starter node.

The problem that we will explore with this report, is if there is a way to optimize the seed-selection of the most influential nodes in a graph. The report will focus on how to use hardware to achieve an optimization to this problem. One idea is to try to use the irregular memory access that is requested during a BFS search.

Chapter 2

Background

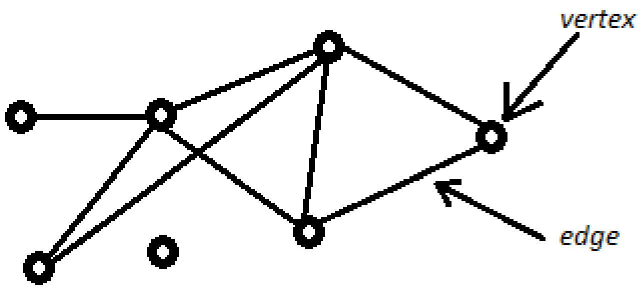
In this chapter, we will look at the fundamental concepts and background information needed to understand the the different infections model, problem with seed-selection for social networks, and solving Breadth-first search using matrix multiplication. This chapter will contain notations that we would use throughout the report and used in the field. One aspect we will focus focus on, is how we can performe graph algorithm such as breathd irst search as matrix multiplication.

2.1 Network terminology and glossary

The fundamental unit in a network is $Vertex(pl.vertices)$ sometimes called a node (computer science). For this report, vertex and node would be used intergangably. The "bridge" or the line connectiong two vertices is called a *Edge*, wich serves as a connection between verices as shown in Figure **??**. Different network have different types of edge, some can be *Directed* or *undirected*. Directed edge is where a edge runs in only one direction (such as a one-way road between two points), and undirected if it runs in both directions. Directed edges, which are sometimes called arcs, can be thought of as sporting arrows indicating their orientation. A graph is directed if all of its edges are directed. An undirected graph can be represented by a directed one having two edges between each pair of connected vertices, one in each direction.

Each node have a value that is known as *Degree*. Degree for a node is the number of edges connected to a vertex. Note that the degree is not necessarily equal to the number of vertices adjacent to a vertex, since there may be more than one edge between any two vertices. In a few recent articles, the degree is referred to as the connectivity of a vertex, but we avoid this usage because the word connectivity already has another meaning in graph theory. A directed graph has both an in-degree and an out-degree for each vertex, which are the numbers of in-coming and out-going edges respectively. The *Component* to which a vertex belongs is that set of vertices that can be reached from it by

Figure 2.1: Simple network



paths running along edges of the graph. In a directed graph a vertex has both an in-component and an out-component, which are the sets of vertices from which the vertex can be reached and which can be reached from it.

A *Geodesicpath* is the shortest path through the network from one vertex to another. Note that there may be and often is more than one geodesic path between two vertices. The *Diameter* of a network, however, is the length (in number of edges) of the longest geodesic path between any two vertices. A few authors have also used this term to mean the average geodesic distance in a graph, although strictly the two quantities are quite distinct

2.2 Network

A *network* is a collection of *Vertices*, commonly known as nodes with *edges* connecting them together[**?**]. The edges serves as a connection, or a "bridge" between the nodes, while the nodes can reprecent something, or containing information. In the real world, multiple systems takes the form of networks around the world, examples are the internettt, the World Wide Web, social media like Facebook, twitter etc. There are different types of network or *graphs*. These include from *socialnetwork*, *sinformationnetworks*, *technologicalnetworks* ,and *biologicalnetwork*. Each of them have a different properties, but we will focus more about social network.

A social network is a set of people connected to each other via some form of contact or interactions[**?**]. The nodes are people while the edges are the connections between peoples. The social network display information regarding connection, interaction or location of a set of people.It forms patterns regarding friendships, business interactions between companies and families history/ ancestral tree. The social network is often used in social science[**?**]. Some noteable experiments are [**?**], which looks at the small world problem. The small world problem can be summarized as: "what is the probability that any two people, selected arbitrarily from a large population, such as that of the United States, will know each other?". [**?**]. This in itselfe is not that interesting, the [**?**] takes that altho person a and z does not know each other, do they have a set of individuals $\{ b_1, b_2, ...b_n\}$ who are mutual friend or even a "chain" of such individual$(a - b - c - ... - y.z)$.

Some properties that networks often exhibits are the small world effect, transitivity or clustering, degree correlations, and community structure[**?**]. Different networks would have some of the specific properties and values. Social network would have a small world effect of around 6 degree, with high transitivity and have a community structure, while a information network, such as a citation network, would most likely have high transitivity since citation and publication have different publication date. Paper A refers to Paper B, which refers to paper C, but paper C can not possibly refer to papaer A since C is clearly published before paper A. [INSErt fIGUre.]

2.2.1 The small world effect

The small world effect was first demonstrated by Stanley Milgra in the 1960s during his famous letter passing experiment[**?**]. The experiment was about passing letters from person to person to reach a designated target with only small steps. For the published case, the chain was around six [**?**]. This shows us that for most pair of vertices in a network can reach each other with a short parth. A more precise wording is that "Networks are said to show the small world effect if the value of l scales logarithmically or slower with the network size for a fixed mean degree.". [**?**]. We have defined l to be the mean geodesic distance between vertex pairs in a network.

For the information diffusion problem, this kind of effect would result in that the iffusion through a network would need around 6 steps to have traveled through the entire network. Meanning that most node can reach each other through a relatively small step.

2.2.2 Transistivity/clustering

Transistivity, also sometimes called clustering is the properties that shows that there is a high number of triangles in the network. A triangle in a network is if node A is connected to node B and node C, and node B and C is also connected to each other, thus creating a triangle. In social network, we say that a friend of your friend is likely to be your friend too[**?**]. The transistivity is often used to show $networkdensity$.

For information diffusion and seed selection, this would mean that picking nodes that are geographically close/neighbour to each other, would have a smaller impact/spread then picking two not neighbouring vertecies. By picking two nodes that are connected to each other, they would likely share a common neighbour thus having a limited reach.

2.2.3 Degree distribution

Degree distribution is the histogram of degrees of vertex. For random graphs, the distribution would most likely be a poiison distribution or binomial. For real world network, the distribution would often have a highly rightskewed distribution. Resulting in that the distribution has a long right tail of values. The probability P_k is the probability that the new vertex v have the degree k.

The degree distribution shows us how the degree to the graph is distributed. For social graph, the degree distribution is often in the shap of [INSERT FOIGURE of degree distribution histogram]. For the information diffusion, this distributon shows us how many high degree nodes there are in the network, those node would likely be high prioritized node and have a large impact.

2.2.4 Degree corrrelations

As [REFER TO FIGURE OF DEGREE DISTRIBUTION] show, the network have a degree distribution with few high degree nodes and multiple low degree nodes. One interesting properties is the degree correlations. Degree correlations is how the high degree and low degree nodes connects to each other. One question is if high degree nodes tends to connect to other high degree nodes, or do they prefer to connect to low degree nodes. It turns out that both incidents is found in networks[**?**]. We can see that for all social network measured in [**?**], the social networks are assortative, meaning vertices have a selective linking, where high degree vertex connects to other high degree vertex[**?**]

For data diffusion, this kind of behaviour would result in wasting a seed by picking majority of high degree node for starting seed, since most of them would be connected to each other. One solution is by mixing the selection, choose some percentage to be high degree, and some with lower degree.

2.2.5 Network resilience

For most network model, there is the need to remove nodes from the network. Removal of a node can have no effect on the network as a whole, or it can be devastating. Network resilience looks at how the network can ressist to such a removal. There are two different removal scheme, the random removal where nodes are randomly picked and removed, or the targeted removal where specific nodes are removed depending on the critteria.

The experiment mentioned in [**?**] by Albert *etal* showed that for a subset of network representing the internett and the world wide web, targeted removal had a larger impact then random removal. The targeted removal removed the highest degree nodes from the network, and the random removal removed nodes randomly. The random removal had a minimal effect on the network, while the targeted removal had a much larger impact, the mean vertex-vertex distance increased. They proposed that the internett was highly resilient to random removal, while much more vuarneble to targeted removal.

There are other studies that proposed a different interpetation about the data found.

In an example of information diffusion, a targeted removal would result in massive change to the diffusion path. By removing high degree node would result in remove influential nodes and limit the spread of the data. Removing important node connection two different communitys would result in isolation and no path towards other community.

2.2.6 community structure

One properties that is often observed in a social network, is the community structure. The community structure is where a group of vertecies having high density of edges with each other, while having lowe density of edges to other "community". We can see an example of the community structure clearly displayed from [**?**]. Where we can see the playground was divided into different community.

This type of community structure would have a lagre impact on how the algorithm would select a seed for information diffusion. If all the seed would be selected in one community, the probability of spreading over to other community would be smaller, them having a seed be in the other community.

2.3 Matrix notations

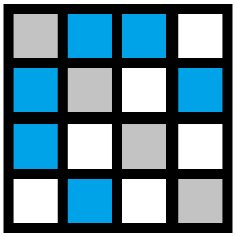
By using a linear algebraic approach to solve graph algorithms often gives us a variety of benefit include easier implementation, higher preformance and syntactic simplicity. [**?**]. There are multiple reasons to do graph algorithms as linear algebra. This can show potentially improvement, potentially optimization and be more visually clear how the algorithm is.

2.3.1 Semiring

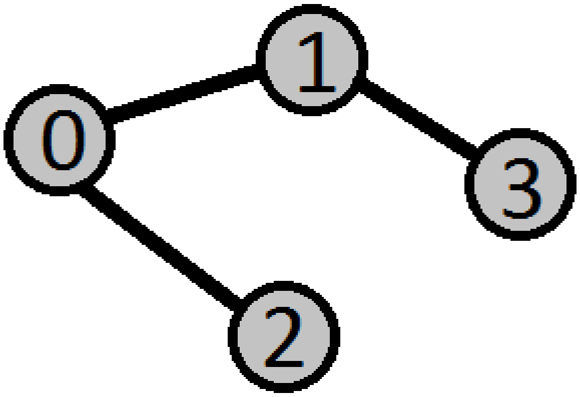
A *semiring* is a set of elements with two binary operations. The two operations are often known as "addition" and "multiplication". By this definition, [SYMBOL FOR BOOLEAN] and [SYMBOL FOR INTEGER]. are both semirings. One way to performe graph algorithm, is apply matrix multiplication over such semirings. One of the common notation in writing the semiring, is $a + b$ and $a * b$. This can be confusin, so often there will be some special semiring operator.

2.3.2 Sparce Matrix

A sparce matrix is an matrix with few non-zero elements in them. The social network can often be reprecented with a sparce matrix. An sparce matrix have



(a) The ajacency matrix



(b) The graph corresponding to the adjacency matrix

its majority of element as zero, as shown in figINSERT FIGURE OF SPARCE MATRIX. We can represent a social graph in the form of a sparse matrix. In an adjacency matrix, a cell containing 1 means a connection is presence, while a 0 is lack no connection. So, as we can see in the Figure **??**, as we can see at coordinate $[1, 3] = 1$. This tells us that there is a edge between vertex 1 and vertex 3. The adjacency matrix is commonly used to represent an graph such as Figure **??** and Figure **??**.

2.4 Data diffusion

Data diffusion is looking at how information is propagated through a network or a graph. An example would be how a new Internet meme, a new trend or how a new disease is spread through a community. The process consist of a set of starter nodes that are "infected", during each time-step, there are a chance that the "infected" node would "infect" its neighbor. To start such a simulation, we would need to first pick out a set of initial "starter node". These k seed nodes is a set of node that in the initial time-step is infected. They will pass on the information/infection during each time-step and the information/infection will propagate through the network.

2.5 Basic Diffusion Models

When we talk about information propagation, we can look at how diseas or technological innovations is spread through a social network. We can simulate those kind of behavior with different diffusion models. There are two basic diffusion models used to simulate the propagation of information through a network[**?**], the $linearthresholdmodel$ and the $independentcascademodel$[**?**].

This process can be simulated by looking at a social network and how information propagates through the network. We look at each node in the graph as a person, they can be either active, or inactive. The activation of a node depends on which diffusion model we choose. A active node is "infected", while the inactive is the "healthy" ones. The activation of each nodes is dependent on which model we pick.

The linear threshold model uses a threshold θ_v between the interval [0,1], which represent the faction of $v's$ neighbors that need to be active to activate node v. The Linear Threshold Model activates the current node v when the weight $b_{v,w} > \theta$ from its neighbor w outweighs the θ_v. This is more in line with the situation were each person have a chance to adopt to a new trend if exposed to the trend enough time by his close friends. An example would be a product promoted on social network like twitter and Facebook. The user will adopt the new trend if he is exposed to it from enough friends or idols.

The independent cascade model changes states with a probability $p_{v,w}$, where v is current node, and w is it's neighbor. During each propagation, the node v have a $p_{v,w}$ chance to change state if the neighbor w changed state. If during time-step t a node v changed state, its neighbor w would have a $p_{v,w}$ chance to change state in the next time step. An example here would be spread of a disease. The current node$v$ have a chance(θ_v) to infect its neighbor

We can se the similarity between the breadth first search and information diffusion. Breath first search, as mentioned before, starts at the root node and add the root nodes children in a queue, then tales the first childe node in the queue and adds all its childen to the queue. Each new child node is added to the back of the queue, while finished node is placed in a $finished_queue$. The sequences of nodes tested is the same as during an information diffusion, where the diffusions neighbour is placed in the $boarder_queue$ and having a probability to be infected. If the node is infected, then that nodes neighbour would be added to the $boarder_queue$, while if the node is not infected, the neighbour to the node would be safe. The information diffusion can therefore be looked at as an breadth first search, with an probability to add the child to the queue.

2.6 Breadth First Search as a matrix multiplication.

By looking at the connectivity matrix, we can see that a connectivity matrix is a graph represented in a matrix format. The BFS can be achieved by looking at the BFS operations as a matrix multiplication [**?**]. If we look at the graph as a connectivity matrix, we can apply matrix multiplication to generate the breadth first search.

Breadth first search is an tree traversal algorithm. BFS start at the root node r, or any arbitary node in a n tree, and stores all its child node in an *queue*. The algorithm then takes the first node from the queue, v_1 and stores all the child node to v_1 in the back of the queue, this process continous until the queue is empty and all the nodes have been iterated over. From the psaudocode we can see:

Algorithm 1 Breadth First Search

```
$dist[\forall v \in V] = -1; currentQ, nextQ = \emptyset$
$step = 0; dist[root] = step$
ENQUEUE(nextQ,root)
while $nextQ \neq \emptyset$ do
    $currentQ = newxtQ; nextQ = \emptyset$
    $step = step + 1$
    while $currentQ \neq \emptyset$ do
        $u$ = DEQUEUE(currentQ)
        for $v \in Adj[u]$ do
            if $dist[v] == -1$ then
                $dist[v] = step$
                ENQUEUE(nextQ, v)
return dist
```

We can also look at how an breadth first seaerch can be executed by appllying sparce matrix multiplication. We can see that teh adjacency matrix is an sparse matrix. We can use sparce matrix umltiplpication to do graph algorithm, one example is the BFS.

2.7 BFS to data diffuison

We can draw an distinctive line between the breadth first search and data diffusion. The breadth first search adds all the childe node from the parent node to a queue, then change the current node to the first node in the queue and repeats until the entire graph is iterated over, or the queue is empty. This is in teorie the same as data diffusion, where information is passed to the connected vertex. The new "infected node" can then pass on the information along to the other nodes that are connected to the new node and so on. This is in practice, the same as breadth first search, minus the searching part.

We can in then draw teh conclution that Independent cascade model, is a modified version of the breadth first search, where each child note have a specific percent to be infected. The itteration is the same as a breadth first approach, but the result of the addition of node is dependent on a random ”coin-toss”.

2.8 greedy algorithm

The greedy algorithm was proposed by Kempe [**?**]. The algorithm is a greedy algorithm where it finds the starter set of nodes S by iterating through all the vertex in V and calculate the total amount of spread. The spread is saved and the k most influential nodes would be choosen.

Algorithm 2 Greedy Algorithm

```
Start with $A = \emptyset$
while $|A| \leq l$ do
    For each node $x$, use repeated sampling to approximate $\sigma(A \cup x)$ to within $(1 \pm \varepsilon)$ with probability $1\delta$
    Add the node with largest estimate for $\sigma(A \cup x)$ to A.
Outpu the set $A$ of nodes.
```

-degree algorithm

Another popular algorithm is the degree algorithm[**?**]. Unlike the greedy algorithm, the degree sort all the node acoring to their degree distribution. The algorithm picks the top k nodes according to the degree distribution.

Random algorithm The last one is the random algorithm. The random algorithm just pick a random starter node.

2.9 Cache oblivious model

A cache oblivious model ignores the cache size and line and designs the algorithm to be cross platform and optimized. An algorithm is *cacheaware* if it contains parameter that can be tuned to optimize the cache complexiity for the cache size[**?**]. Such algorithm have the disadvantage where a adaptation is neede for a new architecture[**?**]. The cache oblivious model is designed for two level of memory, and assumes that tsuch an optimization is optimized for multiple levels as well. There are multiple algorithm designed that shows that such a model actually improves the algorithm. One of such is an cache oblivious sparce matrix multiplication.

2.9.1 Cache oblivious sparse matrix multiplication

Chapter 3

Related works

REltated work, there are multiple people working on this kind of problem. One of the more well known is kemnp, where he tries to solve how to maximize the spread.

Other solution include the hybrid FPGA-CPU arcitecture, which reduces teh computationtime for a bfs over a small world graph by XX%[**?**]

As we can see, there are many different research around this subjects, altho there aren't many that looks at how to accelerate seed selection and the Independent cascade model in hardware, which we will look more at in the end of this report.

Chapter 4

Methode

For this report, we created a python simulation to simulate the data diffusion and the seed selection. We utilized the Pycx libraries to create the GUI, and applied the R-mat generator to create the different network.

4.1 PyCx

Pycx an librarie that help python to generate an GUI[**?**]. The pycx have a clear structure, initialize, observe and update. The initialize part, the graph is generated, the starter seed is found and the position to the graph is generated. The observe part is where python generate graphic for our simulation. for each step, the observe is called to generate a new frame. the update section i called every step, for our program, the diffusion is calculated as each step we can see how the data is diffused. The simulation allowe the information and data to be displayed

4.2 R-mat

One problem during graph analysation and calculation is finding suiteble graphs to analyse. Generate graphs with desirede properties is not easdy to do. One solution proposed by Deeeoayan et al is to use the ”recursive matrix” or R-mat model. The R-mat model generates graph with only a few parameters, the generated graph will naturally have the small world propertie and follows the laws of normal graphs, and have a quick generation speed[**?**]. The R-mat models goal is to generate graphs that matches the degree distribution, exhibits a ” community ” structure and have a small diameter and mathces other criteria.[**?**].

The R-mat generater generates sociall network with the community structure. The different probability for the four partition is :$A = 0.57$,$B = 0.19$,$C = 0.19$,$D = 1 - A - B - C = 0.05$. These different probability was used by

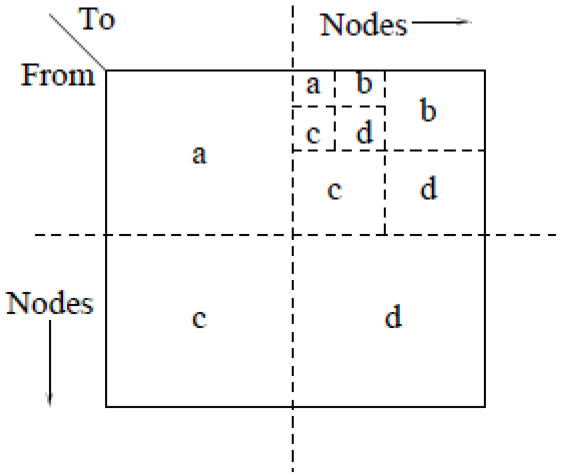


Figure 4.1: The R-mat model

GRAPH500[CITATION NEEDED]. The R-mat generated three different adjacency matrices of different size.

The algorithm to generate such a recursive matrix is as follow: The idea is to partition the adjacency matrix into four equally sized part branded A,B,C,D, like shown in Figure**??**. The adjacency matrix starts by having all element set to 0. Each new edge is ”droped” onto the adjacency matrix. Which section the edge would be placed in, is choosen randomly. Each section have a probability of a, b, c, d, and $a + b + c + d = 1$. After a section is choosen, the partion that was choosen is partitioned again. This continoues until the choosen section is a 1x1 square and the edge is dropped there.

From the algorithm, we can see that the R-mat generator is capable to generate graphs with total numbers of node $V = 2^x$. Since the algorithm partitioned the matrix into four part. This is approach would only generate a directed graph. To generate undirected graph, $b = c$ and the adjacenmcy matrix must make a ”copy flip” on the diagonal elements, like FIgure **??**.

4.3 Adjacency matrices to graphs

For this report, four different sized adjacency matrix was created. One of the restriction to the R-mat generator is that the size of the adjacency matrix have to be 2^n. This resulted in that our adjacency matrix was originaly of the size, $128 \times 128, 512 \times 512 and 1024 \times 1024$. The resulted graphs had multiple singletons and unconnected nodes, for the sake of the simulation, those nodes were discarded and resultede in a graph with 75 nodes and 307 edges, 287 nodes with 2415

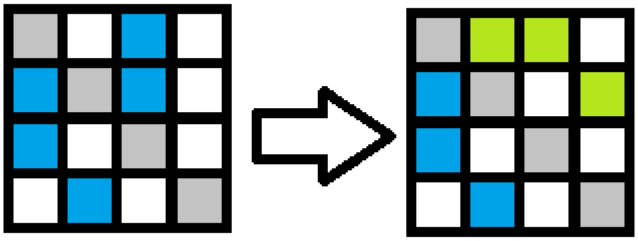


Figure 4.2: How the adacency matrix is fliped on the diagonal

edges and 617 nodes and 8374 edges. This was not supprising considere that for a larger graph, those singletons would most likely result in the outer perifier and small community.

4.4 The algorithm

The simulation creates an social network by reading from the adjacency matrix. We implemented 4 different algorithm. The greedy algorith, the degree algorithm, random algorithm, and a independent greedy algorithm. The algorithm implemented, finds the k vertecies to be the starter nodes. k is [1,2, … 20]. This is to be able to see how teh size of the starting nodes affect the coverage. For each k, the simulation applies the diffusion for 50 times. This is to finde the mean coverage to remove the randomness

The greedy algorithms finds the most influential nodes in relation with the other previous picked seeds. The algorithm starts by finding the most influential nodes s_1 from the entire graph G, in this instance, k=1. The most influential node is found by just choosing a node and see how much spread it will result in, the value is stored with the node, and after the entire G is itterated over, the node with the highest value is choosen. Then the algorithm stores the node in S and applies data diffusion and stores the effect this runs had. The next run, k=2 and the greedy algorithm finds the most influential node s_2 where $s_2 \neq s_1$ and $s_2 + s_1 = maxCoverage$. This is repeated until k=20. The new run, the seed selection will keep the previous selected seed and during the finding maximum coverage phase, the previous choosen seed will have impact on the run.

The degree algorithm chooses the vertex with the highest degree. Unlike the greedy algorithm. The degree algorithm just finds the vertices with the highest degree. The algorithm chooses $s_1, s_2\dot{2}_k$ from G that have the highest degree. One of the problem would be that hihger degree nodes would often be connected to each other, the community struckture that was mentioned in previous section. The degree histogram shows us that there are very few high degree node, while having more low degree nodes.

The random algorithm picks random vertices as the starter nodes. This is the simplest algorithm, where each runs, a new random node is added to the set S. Then the diffusion is applied.

Chapter 5

Result

From the several runs we got the results from all the run. We can see that the greedy algorithm is better then most of the other algorithm. The random algorithm performed worst out of all the other.

Chapter 6

Discussion

One paper showed how implementing a parallalized bfs algorithm on a distributed system reduced the communication times by a factor of 3.5, compared to a common vertex based approach[**?**]. In scientific application, ,graph based computation is pervasived and often data-intensive, this especially for distributed systems.

ISNERT INFORMATION

It would be interesting to see how the parallelized BFS would improve the computation time for our simulation.

Chapter 7

Conclution

As we have seen here the

Bibliography

[1] Magnus Jahre Yaman Umuroglu, Donn Morrison. Hybrid breadth-first search on a single-chip fpga-cpuheterogeneous platform. 2015.

[2] Éva Tardos David Kampe, Jon Klein. Maximizing the spread of influence through a social network. 2003.

[3] Éva Tardos David Kampe, Jon Klein. Maximizing the spread of influence through a social network. 2003.

[4] M. E. J. Newman. *The structure and function of complex networks*. 2003.

[5] Stanley Milgram Jeffrey Travers. An experimental study of the small world problem. 1969.

[6] Stanley Milgram. The small-world problem. 1967.

[7] Stanley Milgram Fjeffrey Travers. An experimental study of the small world problem. 1969.

[8] John Gilbert Jeremy Kepner. Graph algorithms in the language of linear algebra. 2011.

[9] Christos Faloutsosx Deepayan Chakrabartiy, Yiping Zhanz. R-mat: A recursive model for graph mining. 2004.

[10] Kamesh Madduri Aydib Buluc. Parallel breadth-first search on distributed memory systems. 2011.