UNIVERSITY OF OSLO
Department of Informatics

Social graphs - change of metrics' distribution with scale

Master thesis, 60 credits

Olga Voronkova

January 1, 2014

Abstract

Social networks are systems that are generally composed of multiple entities interacting with each other to provide a desired functionality. The interactions between these entities can be modeled as graphs. Presenting these interactions in terms of graph models allows system designers to not only investigate and reason about their systems but also to design new solutions and applications.

Real interaction data is required to build graph models. However, in many scenarios it is difficult to obtain real data because of restrictions, such as privacy issues, scale of the system and administrative restrictions. There has been done a great amount of work in the social graph crawling and modeling field, however there has not yet been conducted a study of how different metrics behave when the graph size is changing, combining observations from both modeling and sampling.

Our contribution with this work is mapping how degree, clustering coefficient, closeness and betweenness distributions are affected by scale both for Watts-Strogatz models and when sampling with Random Walk, Breadth First Search and Metropolits-Hasting Random Walk. We argue that clustering coefficient distribution gets further away from the original values for smaller graphs, and that the rest of the metrics are not affected by scale. We also show that joint degree distribution metric is not under control of Watts-Strogatz model.

Key words: Social Graph Sampling, Graph Modeling, Scale, Metrics

Preface

This thesis is the result of a 60 point mastert project completed at the Programming and Network research group at University of Oslo, Department of Informatics.

I would like to extend my gratitude to my advisors Roman Vitenberg, Abhishek Singh and Frank Eliassen for all the help with the code, the writing process, their feedback and fruitfull discussions - I couldn't have made it without their support and guidance.

During the past half year I shared workspace with PhD students, where I got to meet great people. I would like to thank Vinay Setty, Lucas Luiz Provensi and Navneet Kumar Pandey for their support, cookies, great food and Xbox bowling!

I would also like to thank Bjørn Einar Bjartnes and Harald Solstad Fianbakken for proof reading my thesis and making me rewrite the parts not understandable to anyone, myself included.

Warm thanks go to Tom for all support and home-cooked meals and my mom for being the cheerleader team i needed.

And also, to spider solitaire for keeping my sanity.

Nomenclature

BFS Breadth-first search

CC Clustering coefficient

JDD Joint degree distribution

K-S Kolmogorov Smirnov

MHRW Metropolits-Hasting random walk

OSN Online Social Network

RW Random walk

RWRW Re-weighted random walk

Contents

List of Figures

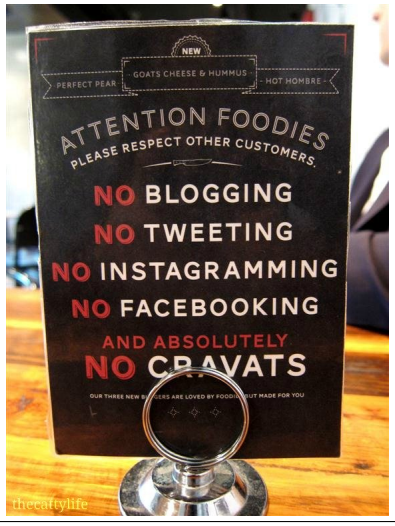
List of Tables

Chapter 1

Introduction

Online social networks (OSNs) are a modern phenomena that has exploded in popularity the past years, with constantly new social sites popping up and the old ones growing in range. The reason for this massive popularity is simple - social networks are entertaining! They introduce a whole new lifestyle, where you can either take an active part by writing blogs, posting updates, pictures, marking your favourite movies, posting pictures of your food (in fact, some of the restaurants - like the one illustrated in Figure 1.1, have begun to ban custommers who apply any use of social media on their food).

Figure 1.1 "When Food Instagramming gets Out of Control" (borrowed from [MessyNessychic, 2013])



There are numerous online social networks with different focus, such as networking, dating, debating, knowledge sharing, travelling and more. The most visited of them, according to [Google, 2011] is Facebook. Facebook de-

scribe themselves as a social platform that " helps you connect and share with the people in your life" (borrowed from [Facebook, 2013a]). Even though primarily this is a network for connection, Facebook has many other features, making it appealing for large crowd. With Facebook tools you can create events, manage groups, send messages, chat, push photos and videos, sign into a location, poke somebody, have video chats and so much more. Social networks are entertaining. They are so intertaining that they become addictive. Facebook has indeed become an important part of everyday life, with millions (latest statistics from [Brain, 2013]) of users checking their Facebook on daily basis - also popularly known as "the Facebook addiction", which has in turn lead to numerous jokes, such as the image in Figure 1.2.

Figure 1.2 The Facebook-addiction effect (borrowed from [Risen Sources, 2011])

1.1 Social network research

Social networks are an object of study in many fields such as anthropology, biology, economics, geography, information science, organizational studies, social psychology, sociolinguistics and a number of other scientific and not-so-scientific groups.

Studying social graphs can for example show how a disease will spread amongst a population. For instance influensa virus is very contageous and will therefore spread fast through the neighbors of neighbors further to their neighbors - a typical breadth first search algorithm.

Commersial business have also started to take advantage of the potential of social networks as well targeting users by their friend circles, interests and gender.

Recently there has been a big case in the newspapers after Edward Snowden's leakages of top secret government documents, revealing that National Security Agency is monitoring US citizens through their surveillance program. The Prism program ([Guardian, 2013]) was tapping directly into user data collected from the servers of social sites such as Google, Skype, YouTube and Facebook.

Whom we have in our social networks apparently also control our mood. Studies have shown that if a person is happy, his or her friends have a 25%

higher chance of being happy (described in [Passmore, 2011]).

Social networks are good for spreading ideas and receiving inspiration for new ideas. Networks with so called "weak ties", or bridges between clusters of user groups are therefore more usefull for their users than tightly connected networks where everybody knows everyone, as these users are more likely to share same interests and views and therefore have less new ideas to share with each other. On a sidenote, there exists a theory by Richard Dawkins ([Dawkins, 2006]) where ideas are using humans to populate, spread and grow. In this case, human social networks play a big role in the evolution of ideas.

Even though we usually associate social networking with leisure activities, they have a great effect on our carees as well by creating informal connections not only with our coworkers, but also employers, potential emloyers and people working in the same field. How problems are solved, the structure of organizations, whether or not individuals manage to achieve their goals, job seeking and many more areas within our careers can all be a part of our social networking. There have been built several social sites specifically aimed at career networking, such as LinkedIn. Many workplaces also encourage their employees to actively take part in programming forums such as StackOverflow, creating their own blogs, taking parts in conferences, publishing papers and simply chatting with coworkers on local social sites such as Yammer.

Social network analysis reveals networks' impact on Internet traffic. The results of such analysis can help improve robustness and security of social networks, as well as mapping the effect they may have on Internet in the future.

Facebook has recently launched a Beta version of Social Graph Search (January 15th, 2013, [Facebook, 2013b]), making an interesting case of what can be done with social network analysis. Facebook graph search results are unique for each user, and include what your friends have shared with you on the topic you are searching for (for instance which restaurants in Oslo were recently visited by your friends who like rock music), and other information that is publically available on the topic. From 1st of October 2013 it is also possible to search for status updates your friends have shared. This feature can be very usefull for creating a communal feeling, whenever you need to borrow something, or want to know who of your other friends would like to see the same movie noone you have spoken to wants to see, or if you need someone to rent a place to, you can search for "posts of my friends who <talk about the same thing as me>""

1.2 Scale and properties of social graphs

The previous section has shown how important social network analysis is for many fields. Such analysis is however difficult to perform in full scale, since the networks have grown enourmously over the past years. Running a simple algorithm on a full scale graph may turn out to be an impossible or very time consuming task.

Instead, it can be very usefull to take a smaller sample or a model of a graph, that posesses the same properties (the distribution of metrics) of the original graph. By verifying that the algorithm works on the smaller scale graph, we can predict the output of the same algorithm on the original full scale graph. For instance, if the desired use of a social graph is to see how an epidemics will spread among the population, we can extract a smaller portion of the population - a so called representative sample, and run the algorithm that will simulate the disease spreading. From the measurements it is then possible to calculate how the same algorithm will behave on a bigger scale.

There is unfortunately no single answer up to date for which metrics to use for each individual experiment. The testers should always question which properties of the graph their tests will affect, and adjust the metrics preserved with scale consequently. It is however challenging to locate the exact properties that should be preserved for each experiment. The challenge also lies in replicating a graph in smaller size so that all of the desired properties are preserved.

By studying the structure of social networks researchers attempt to replicate the desired properties (metrics) of the original graph on a smaller scale. In order to achieve a smaller replica of a large graph, various techniques such as sampling and modelling have emerged. There however still is no single commonly accepted solution, leaving many possible trails to explore in this field.

1.3 Problem definition

One of the main issues with graph crawling algorithms today is that they may produce graphs with significantly different topological properties from the original graph. This can cause incompleteness of data and introduce a bias to the sample - for instance it is in the nature of Breadth First Search algorithm to oversample nodes with high degree, by traversing all neighbors of a node before proceeding to the next level neighbors of neighbors (see Section 2.3). It is therefore important for result precision to correct the bias before analyzing the sampled graph. There are several ways to correct bias, either with modifying the crawling algorithm or modifying the resulting graph after the algorithm was run. Before correcting the bias however, it is

essential to locate which algorithm produces which bias, and which metrics this bias in turn will affect. There is no straight-forward way to do that yet, other than by running the algorithm and studying metrics of interest. Metrics most commonly measured when trying to estimate the algorithms' bias are Node Degree and Clustering Coefficient. Other metrics however have not been given as much attention to.

Another technique of producing a smaller graph is modelling. The problem with modeling is that there is no way to control all the desired metrics - models are built to take into consideration a handfull of metrics, giving no guarantee for how the rest of the metrics will behave when scale for modeled graph changes. Often model documentation only specifies a few of the metrics that are controled by that model, leaving the rest for the reader/developer to find out.

1.4 Research questions

Taking into consideration problems with both sampling and modeling techniques, we wish to investigate further and compare how metrics will behave on graphs produced using both of the techniques, while we modify the size of the graph. Our research is guided by the following questions:

1. "How do graph metrics change with scale?"
2. "Which sampling algorithm will perform better in regards to preserving metrics when scaling?"

In order to answer our research questions, we need to find out which metrics are controled by the model we are using, meaning which metrics we can safely use for comparison, knowing that the values we get from our tests are representative to any other graph built from the same model using the same parameters.

Our intuition is that metrics distribution will get further away from the original graph (become more imprecise) when graphs get smaller, since it would be more difficult to preserve the original properties when there are less nodes.

1.5 Scope of the thesis

Given that our research questions are wide and can lead to numerous interpretations and possible trails, it is important to find a scope that will be reasonable to achieve within the time constrains of a master thesis. We will look at the metrics' behavior both for the samples we obtain by different algorithms, as well as for constructed models to identify patterns in the behavior. We will only study samples of our Facebook BFS graph and model

generated graph of equal size, and use one of the known models (Watts-Strogatz) to determine the "default behaviour" of metrics. It would have been interesting to compare different real Facebook samples, however these are difficult to obtain due to Facebooks changed policies, a problem discribed in more detail by [Gjoka et al., 2010].

We have compared graphs based on degree, clustering coefficient, joint degree distribution, closeness and betweenness metrics, however there are many more metrics that could have been included, such as k-connectivity, 3K, likelihood, sampling algorithm's start node and more.

Is it safe to assume that the distribution of metrics in a sampled or modelled graph will correspond to the original graph? Will the distribution of metrics be different for different scales, but the metrics will still follow a scaling rule? Will such scaling rule be usefull to researchers studying social graphs? These questions are very important but also difficult to answer. This work brings us one step further by taking a closer look at sampling and modeling and how the graph scale affects some of the metrics chosen.

1.6 Research approach

The aim of this thesis is to investigate how metrics behave when we reduce the size of the graph. We apply both sampling and modeling techniques to achieve smaller graphs, and then compare how the results vary for sampled and modeled graphs. For experiments with graphs it is common to use the term "ground truth" to describe the baseline for comparison - the original graph that is either being sampled or modeled. For this project we operate with two different ground truths to run sampling algorithms on - a Bredth-First-Search (BFS) sample with 10 000 nodes from Facebook data, and a graph constructed based on Watts-Strogatz model, equal in size to the BFS sample. The graphs we are working with are built of nodes that represent users of the social graph, with links that show the users' connections in terms of friendship.

Metrics describe structure and properties of a graph. By comparing metrics' distribution for two graphs it is possible to say how well they resemble each other. Since the most commonly used metrics are clustering coefficient and node degree, we wish to expand the experiments to include a wider range of metrics. The metrics we are studying in this thesis are degree (how many friends a user has), clustering coefficient (how close friends of a user are to forming a clique), joint degree distribution (how triples of nodes of certain degree are connected to each other), betweenness (amount of shortest paths passing trough a node) and closeness (average distance from a node to all other nodes).

Our experiments can be grouped into three steps:

1. Sampling

The task first consists of running sampling experiments on our two ground truths - BFS sample and Watts-Strogats graph. The sampling algorithms applied are Random Walk (RW), Metropolis-Hastings Random Walk (MHRW) and Breadth-First Search (BFS). Sampling algorithms collect nodes "as they go", depending on the algorithms specifications, untill a given amount of nodes has been reached. At the end of our sampling experiments we filter out the one algorithm that has the closest results to the ground truth in terms of how metrics change with scale.

2. Modelling
   The second step is to construct models of different size and analyze the difference in metrics' change with scale between the modelled ground truth and the smaller models corresponding in size to our samples (5000, 2500 and 1000 nodes). Modelling a graph is imitating some of the desired properties (for instance how many friends an average user has), the nodes and links are constructed based on the parameters given to the model, and the properties that are known to be controled by the model are therefore more predictable than in a sampled graph.

3. Sampling vs Modelling
   The third step is to compare the metrics' behaviour with scale between samples and models.

1.7 Results

We contribute to the social graph research field with the following observations, further described in Section 5.3:

1. Clustering coefficient was affected by scale when measured on models as well as on some of the samples.

2. Joint degree distribution was not under control of Watts-Strogatz model.

3. When constructing Watts-Strogatz model, clustering coefficient was affected by the randomness parameter.

4. The degree parameter for Watts-Strogatz generated graph became a restriction during sampling, preverting a stable distribution of closeness, betweenness and degree metrics in the sampled graphs.

5. Watts-Strogatz Small World model did not preserve all of the important properties of social graphs (the correlation between clustering coefficient and degree metrics).

6. MHRW and BFS stood out when preserving metrics with changing scale.

7. Real graph samples can be difficult to immitate with Watts-Strogatz model, especially in terms of clustering coefficient, closeness and betweenness metrics.

1.8 Chapter presentation

This thesis is organized into five Chapters. Following "Introduction" is the "Background" Chapter describing state of the art for social networks research, as well as related work. The next Chapter, "Samples, Models, and their Analysis" presents our methodology, describing the approach used for constructing, running and evaluating the graphs. The "Testing and Evaluation" Chapter shows metric measurements we extracted from comparing our BFS "ground truth" to crawled samples as well as to model ground truth and model-generated graphs of different sizes. The Chapter presents the preferred algorithm to use for sampling, as well as an overview of which metrics are affected by scale and which are not. The "Conclusion" sums up our experiment and sketches some directions for future work.

Chapter 2

Background

While Chapter 1 presented the problem we wish to address, this Chapter aims to give a general description of state of the art related to this work. Unfortunately for an individual work (but fortunately for research of course), there exists a huge number of articles on social graphs and their properties, making it impossible to include it all in a scope of one thesis. The articles we cite are the main representatives in the field that we wish to investigate further, however it is important to note that the state of the art presented in this Chapter is only a tip of an iceberg of the existing litterature.

First we describe the background for the type of data we are working on - social networks, followed by a short description of social graphs in Section 2.1. Social graphs have specific properties, such as scale free, power law, dense core, small world, and the correlation between degree and clustering coefficient metrics. In this thesis we only look deeper into cc-degree correlation, however scale free, power law, dense core and small world properties are important for a deeper understanding of social graphs. The aforementioned graph properties, together with metrics that we have used for comparison of the sampled or modeled graph to the original graph are described in Section 2.2. Metrics that we look at in this thesis are degree, closeness, clustering coefficient, betweenness and joint degree distribution. These are the most common metrics applied in social graph research, however there are plenty more to look into and compare (described in [Passmore, 2011]). Further the Section describes tools and techniques we have used for analyzing our graphs. Gephi is used in this thesis as a backup tool to validate the correctness of our final results for degree, closeness, betweenness and clustering coefficient metrics. Kolmogorov-Smirnov test is used as a measure of distance between two distributions. Pearson correlation coefficient calculates the correlation between clustering coefficient and degree metrics. Standard deviation was used to map how metrics' measurements varied if we ran several identical runs, to check whether the results we achieved were representative for each graph.

The next Sections 2.3, 2.4 and 2.5 present different types of approaches for minimizing the graphs while keeping their main properties - graph crawling, graph modeling and the combination of the two. These are the three main alternatives researches choose from when creating a graph of a given size. We have chosen to look at modeling and sampling separately, focusing on the comparison between the two techniques.

For graph sampling we chose Random Walk, Metropolitan Hastings Random Walk and Breadth First Search algorithms, based on the algorithms researched on in related work we have looked at. There are however other algorithms to choose from, such as Forest Fire, Snowball sampling, Debth First Search and many more. Graph Modeling . Alternatively we could have chosen Nearest Neighbor (described in [Sala et al., 2010] as the best model to depict a social graph), Kronecker Graphs, DK graphs, or one of the other models.

2.1 Online Social Networks

Social networks are social structures of multiple entities (users) connected by friendship, common interests and other bonds, as defined by [Passmore, 2011]. In social networks, the interaction between users are more important for the analysis than the attributes of individual users. The interactions between the users can be modeled as graphs. This allows system designers to not only investigate and reason about their systems, but also to design new solutions and applications.

2.1.1 Social Graph

A social graph is a mapping abstraction of social network, where individuals are abstracted to nodes, and their relationships to links. The graph is usually highly dynamic, and therefore hard to traverse completely by crawling, an issue addressed by Tad Miller in his blog [Miller, 2010].

2.2 Analysis of online social networks

In this Section we present metrics we chose to measure in our experiments, describe some of the most common properties of social graphs and introduce the analysis tools that were used to measure and compare metrics and properties of our graphs.

2.2.1 Metrics applied in this thesis

[Mahadevan et al., 2006b] stated that up to date there exists no systematic way to determine which metric to use in a given scenario. Metrics are therefore manually picked out for each case. In ours, we have looked at the metrics

we considered the most essential to graph analysis, based on the metrics most frequently found in related work.

dK

[Mahadevan et al., 2006b] present a methodology for comparing graph topologies - dK-series of properties, that reflects how groups of d amount of nodes with given degrees interconnect. dK-series is the simplest basis for statistical analysis of correlations in a complex network. dK-series presented in the article support inclusion and conversion requirements. Inclusion implies that all properties for series with a lower d, are satisfied by the dK. Convergence means that for a large enough value for d, the generated graph will become isomorphic (identical based on given properties) to the original graph. Therefore the authors point out that any metric defined on the original graph will eventually be captured by dK-series, with a large enough d. However it should also be taken into consideration that with a larger d, the amount of probability distributions also increase drastically, while only one of them is isomorphic to the original graph. 0K graph lacks high degree nodes, 1K has high degree nodes compressed at the core, 2K graph is pushing the high degree nodes further to the perifery, while 3K topology is similar to the original topology. The authors conclude that 2 is a sufficient enough value for d for most practical purposes, and that 3K-series result in an almost identical graph to the original for all Internet-like graphs.

The following is a brief summary of the three first metrics of DK-series:

1. 0K-graphs - average node degree
2. 1K-graphs - node degree distribution (probability calculation of nodes having degree k. This is reflected in Facebook through amount of friends a user has)
3. 2K-graphs - joint degree distribution (number of edges connecting two nodes/interconnectivity, or number of nodes connecting to other nodes of different degree)
4. 3K-graphs - interconnectivity among triples of nodes

A problem with dK-series is that it doesn't support non-integer values for d. In cases where a property is not captured by k value of d, while it is overrepresented in k+1 value of d, it will overconstrain the algorithm.

The higher the value of d, the more of different metrics are covered, making the dK-metric the best metric to use, however also the most complicated and time-consumig one to measure. This is why we have chosen to focus only on 1K and 2K metrics.

Node degree is perhaps the most ground metric of them all. For graph analysis it is an important property that identifies the "key players" in a network. Removing these nodes may cause the partitioning of a whole network. Degrees tell us also about the structure of a graph, whether a graph is random (random number of neighbors for each node), or if there is a pattern, as explained in [Steen, 2010].

Clustering Coefficient (CC)

Clustering coefficient, also known as transitivity, shows how close neighbors of the average k-degree node are to form a clique. A clique between neighbors of a node means that every node is connected to all other nodes.

This metric definition was borrowed as well from [Mahadevan et al., 2006b]. [Gjoka et al., 2010] describe it also as the "relative number of connections between the nearest neighbors" of a node.

For practical purposes, clustering coefficient can be used for identifying communities. Members of a community tend to be tightly connected to one another, while there are only few connections between two different communities. Each user however can be a member of several communities. Mapping clustering coefficient properties for nodes in a network is important for instance for information spreading, as highly clustered networks are slower on algorithms for gossiping (epidemics). For more detailed explanation on implementation and appliance of clustering coefficient, see [Steen, 2010].

Betweenness centrality

Betweenness centrality is also known as shortest-path betweenness and measures amount of shortest paths that traverse a node. This metric definition was borrowed from [Ducruet and Rodrigue, 2013]. Another way of defining betweenness (by [wiki.gephi.org, 2013b]) is that it is a measure of how often a node appears on shortest paths between nodes in the network. A high betweenness centrality can be interpreted as the user that connects different parts of the network, according to [Hirst, 2010]. A user with high betweenness centrality is a popular user in a social network, and probably in real life as well. Shortest path centrality can be usefull during construction of subway stations for instance. The architects should determine which stations are most likely to have most visitors (typically central or crossroad stations), and therefore expanding those stations to have greater capacity of travellers per day.

Closeness centrality

Closeness centrality of a node is described in Networkx api ([NetworkX, 2013b]) as "the reciprocal of the sum of the shortest path distances from that

node to all other nodes. Since the sum of distances depends on the number of nodes in the graph, closeness is normalized by the sum of minimum possible distances." The higher the values of closeness, the higher is the centrality. Or in other words (by [wiki.gephi.org, 2013a]), it is the "average distance from a given node to all other nodes in the network". Returning to our previous subway example, a centralized system should allow users to be able to travel from any point in the city to any other point in a relatively equal amount of time. Here is where centrality plays a role by determining the average amount of stations that should be placed between the main stations, giving a hint of how many different lines and directions that need to be added to the subway net, facilitating users to get around faster and more conveniently.

The above are the metrics we have decided to apply for this thesis. Node degree and clustering coefficient were chosen as a the most common metrics measured, and therefore easy to compare to previously done work. Joint degree distribution would give us a picture of how well connected the graph is, and closeness and betweenness would describe how centrality is spread among different nodes in our graphs. Metrics were mainly chosen by "commonly used and easy to compute" criterias.

2.2.2 Common properties of social graphs

[Mislove et al., 2007] examines multiple online social networks at scale, and is similar to [Gjoka et al., 2010] in that the authors use crawling techniques to obtain data sets from social graphs. According to [Ye et al., 2010], their study is the largest OSN crawling study up to date, and shows that OSNs have "power-law", "small-world" (average path length between two nodes is 6 hops) and "scale-free" properties.

[Yoon et al., 2007] create a scale-free network model and let a random walker traverse it. Based on measuring degree distribution, degree-degree correlation and clustering coefficient, they find that their sampled networks keep the original graph properties such as power law, degree correlation and modular structure, measured through clustering coefficient.

We have chosen to describe deeper some of the properties we considered essential for understanding structure of online social networks, however we will only look closer at correlation between clustering coefficient and degree in this work.

Scale free and Power law

A network is scale-free if it's degree distribution follows a power-law, where high degree nodes tend to connect to other high degree nodes,

and low-degree nodes connect to other low-degree nodes. This is the case for social graphs, as stated by[Mislove et al., 2007].

Power law degree distribution results in a graph with few high-degree nodes and many low-degree nodes. The number of nodes with a high degree decreases exponentially, and the probability of a node having degree k is proportional to the scaling exponent $(1/k)^{\alpha}$, where $2 < \alpha < 3$.

Another typical property for scale free network is that degree distribution remains the same regardless of the range of the graph (for instance taking different ranges of node id's in a graph and plotting them will produce the same visual effect). The power law and degree distribution properties are described in more debth by [Steen, 2010].

According to [Mislove et al., 2007], a scale-free network is recognized by significant clustering among low-degree nodes.

[Leskovec and Faloutsos, 2006] remarked however that in practice social networks will deviate from the power-law property.

Dense core

[Mislove et al., 2007] measured link-degree correlations, joint-degree distribution and scale-free metrics. Their results indicated that the topology consisted of a core of between 1% and 10% of the highest-degree nodes linking to strong clusters of low-degree nodes at the fringes of the network. The core of high-degree nodes are critical for networks' connectivity and removing it would cause a complete disconnection of the graph, making it vulnerable to malicious attacks.

[Steen, 2010] however writes that random networks (meaning also social graphs) consist of a single large component (dense core), with few small components - many of which contain only one node.

The above two observations, agree on that a dense core, of small size, is present in any social network graph.

Small world

By definition, Small-world networks have small diameter and high clustering. [Mislove et al., 2007] have measured both diameter and clustering for the four major OSNs at that time - Orkut, Flickr, Youtube and LiveJournal, and concluded with that social networks obey this property. Small-world effect means that most nodes are connected, and can reach each other in small number of hops through connecting links. [Steen, 2010] describe such paths between two nodes as weak links. In order to disconnect a real-world social graph, it is often necessary to remove 70-80% of nodes.

Correlation between clustering coefficient and degree

[Yoon et al., 2007] discovered that clustering coefficient for a node with degree k was decreasing with increasing k. This has also been observed by [Mislove et al., 2007], where the authors measured that clustering coefficient is inversely proportional to degree.

2.2.3 Tools and Techniques for Graph Analysis

There are numerous tools and techniques for different purposes of graph analysis. For our analysis we have applied Gephi tool, and Kolmogorov-Smirnov, Pearson correlation coefficient, test and Standard deviation tests.

Gephi

During our research we encountered a usefull tool, Gephi [Gephi, 2013], an open source graph visualization and analysis software. Gephi takes an edgelist as input, creates a visual graph and computes many of the most popular metrics. We were able to calculate degree, closeness, betweenness and clustering coefficient, while the only metric not covered by Gephi was joint degree distribution.

Kolmogorov-Smirnov

[Leskovec and Faloutsos, 2006] use Kolmogorov-Smirnov (K-S) statistics value to compare results for graph measurements. Kolmogorov-Smirnov test compares the two samples and indicates how different they are from each other, measuring the furthest point of the two samples. The higher the value, the further they are from each other, while a small value indicates that the two samples have the same distribution.

Pearson correlation coefficient

Pearson Correlation is described in [StatSoft, 2013] as a way to determine whether or not two variables are "proportional", or linearly related to each other. The return value of 0 would imply no correlation between the two datasets, while 1 or -1 would indicate an exact linear correlation. If result is positive, it should be interpreted that both x and y are increasing. A negative correlation means an increasing x and decreasing y.

Standard deviation

Standard deviation is used to find abnormalities from the expected result. The higher the deviation value, the more spread apart is the data, according to [Investopedia, 2013].

2.2.4 Conclusion of Social Graph Analysis State of the Art

In this thesis we will look closer at joint degree distribution (2K), clustering coefficient, betweenness and closeness centrality. We noticed that there was

a higher occurence of these metrics in related work and chose to focus on these metrics in order to connect our experiments to state of the art, to facilitate the comparison and use of our results in future work.

Social graphs have many specific properties. Since we already had measurements both for degree and clustering coefficient, it was natural to look closer into the social graph property of the correlation between clustering coefficient and degree metrics, and how well this correlation is maintained when the size of the graph changes, both for crawled and modeled graphs. The rest of the properties described in this section were not included due to the time constrains.

2.3 Graph Crawling

A problem with OSNs, pointed out by [Mislove et al., 2007] is that since OSNs are getting bigger and bigger, it is difficult or sometimes impossible to traverse the whole graph. Sampling or crawling a part of the graph instead is an inexpensive and efficient solution, and there are a number of algorithms commonly applied for that purpose. Crawling is defined by [Gjoka et al., 2010] as a technique to traverse graphs by visiting a node and then it's neighbors, in the order specified by the algorithm. In different academic papers sampling in our experience can be refered to as a synonym to crawling, as in paper by [Leskovec and Faloutsos, 2006], or random querying for node id's, as for instance in [Gjoka et al., 2010].

2.3.1 BFS, RW, MHRW and RWRW vs uniform sampling

One of the most relevant works for this project is [Gjoka et al., 2010], where authors compare different approaches to crawling OSN graphs - Breadth First Search (BFS), Random Walk (RW), Metropolits-Hasting random walk (MHRW), and Re-weighted random walk (RWRW). They focus on "node degree distribution" metric and other comparison criterias to measure convergence, such as "sizes of geographical network" and "userID space". For parameters they used burn-in rate, total running time (walk length) and thinning (sampling rate). In addition, they compare their algorithms to a sample obtained through UserID rejection (UNI), which they use as a "ground truth". They argue that even though a uniform sample is retrieved by randomly generating user id's and sampling the ones that exist, this approach would not work with user id's longer than 32bits. Authors therefore emphasize crawling as the correct sampling technique for OSNs.

2.3.2 Multigraph crawling

The work by [Gjoka et al., 2011] is the first in line to sample OSNs by combining multiple relationships. Instead of sampling from a graph con-

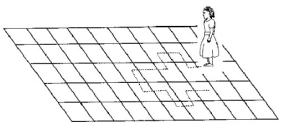
nected through a single social relation (friendship), [Gjoka et al., 2011] create a union multigraph based on connections through memberships in shared groups and events as well as friendships. By applying re-weighted random walk, they crawl Last.fm graph (fragmented, multigraph structured) by different type of relationships separately. The results are then combined into a multigraph, by pairing the relationships into sets (Friends-Events, Friends-Events-Groups, Friends-Events-Groups-Neighbors). The combined union multigraph is then crawled, and results are compared to the ground truth - a UNI sample, collected in a similar way as in their previous work, described in [Gjoka et al., 2010]. The authors show that multigraph sampling improves graph coverage when there are many isolated users without any direct social ties - the connections otherwise not reachable by single-graph sampling.

As mentioned previously, there are numerous crawling techniques for obtaining a graspable sample that maintains the characteristics of the original graph. There are several algorithms commonly used for graph sampling, but for this thesis we have chosen to focus on only three, namely Random Walk (RW), Breadth-First Sampling (BFS), and Metropolis-Hasting Random Walk (MHRW).

2.3.3 Random Walk

In RW, the next move is chosen randomly from the current nodes' neighbors, and will therefore lead to more frequent visits to nodes with higher indegrees. There are however different random walk variations that according to [Gjoka et al., 2010] correct the bias, such as Metropolis-Hastings Random Walk and Re-weighted Random Walk. The mechanism of this algorithm is illustrated in Figure 2.2. In experiments done by [Leskovec and Faloutsos, 2006] Random walk has shown to perform better than Forest Fire, Random Node and Random PageRank Node when decreasing the allowed sample below 50% of the original graph size and down to about 15%.

Figure 2.1 Illustration of Random Walk, where nodes are visited in a randomly fashion from the current node. Image borrowed from [Sprott, 2013]

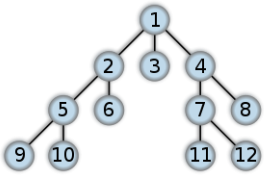


2.3.4 Breadth-first search

Breadth-fist search is one of the more common techniques for sampling, partly due to its ability to collect a full view of a graph region, and partly

due to being quite straigh-forward to understand and use. BFS has shown to be outperformed by MHRW and RWRW in the work by [Gjoka et al., 2010]. The mechanism of this algorithm is illustrated in Figure 2.2.

Figure 2.2 Illustration of Bredth-First Search, where nodes are numbered in the order they are visited. All children of a node are visited first, then followed by children of children. Image borrowed from [Wikipedia, 2013b]



2.3.5 Metropolis-Hastings Random Walk

According to [Gjoka et al., 2010] Metropolis-Hastings Random Walk corrects the weight during the crawl by only moving to a neighbor node if it has lower degree than the current. In experiments done by [Lee et al., 2006] MHRW has shown to produce unbiased samples of undirected social graphs and to perform better in tightly connected graphs, keeping the degree distribution. [Wang et al., 2011] point out that average clustering coefficient varies with the datasets.

[Gjoka et al., 2010] calculate MHRW probability by generating a random number between 0 and 1. The next step in the MHRW walk is permitted only if the quotient of current node's degree and neighbor's node degree is greater than the random number. In their implementation, the rule of "moving only to lower degree nodes" is not strictly followed, and the move to a higher degree node will be allowed if the random number is sufficiently low. This makes it possible to avoid getting stuck at a low-degree node with only high degree neighbors.

2.3.6 Challenges with crawling algorithms

There are several known challenges assosiated with the crawling technique - namely bias, performance, defining a "good enough" sample, and obtaining access to the real graph.

Bias

[Lee et al., 2006] measure degree and betweenness centrality distribution, average path length, assortativity, and CC for performance of node sampling, link sampling and snowball sampling methods mainly for scale-free graphs. They point out that bias of a sampling algorithm can be predicted when

looking for measuring a metric of choice, and investigate bias for each of the algorithms.

Different algorithms introduce different bias - for instance snowball sampling algorithm tends to underestimate the degree exponent of degree distribution property, and produce graphs that are more disassortative than the original. Bias to some metrics can also depend on the network, as is the case for clustering coefficient metric for snowball sampling described in [Lee et al., 2006] [Gjoka et al., 2010]point out that in general BFS and RW are known to create a bias towards higher-degree nodes.

Node Degree Bias

Some algorithms, for instance BFS and RW in work of [Gjoka et al., 2010] and [Lee et al., 2006], tend to pick higher degree nodes, overestimating the node degree distribution metric. The worst-case scenario of this bias would be when the seed node is a higher or equal degree than the amount of nodes in the sample. In that case the sample will be presented as a star network consisting of N nodes, where 1 node is of degree N, while N-1 nodes will have a degree 1, thus giving a different network topology than the original graph.

Collecting a larger sample of nodes from the original graph will correct the node degree bias. To avoid increasing the sample, it is also possible to take a sample of the crawled subgraph in order to fix the degree bias while maintaining crawled sample small, as was done by [Ye et al., 2010].

Clustering Coefficient Bias

Oversampling nodes with higher degree leads in turn to underestimation of clustering coefficient. [Ye et al., 2010] observed that clustering coefficient increases with the size of the sample.

BFS is again one of the good examples where clustering coefficient bias occurs. Even though the overall clustering coefficient is preserved by BFS due to bias, as was pointed out by [Mislove et al., 2007], the algorithm obtains larger average clustering coefficient than the original graph, since CC is strongly dependent on node degree, an observation made by [Wang et al., 2011].

Power-law Bias

BFS tends also to underestimate the level of power-law coefficient, as noted by [Mislove et al., 2007].

Performance

There is also room for improvement in crawling algorithms known today when it comes to performance. It has been shown by [Wang et al., 2011] for

instance that graph properties, such as connectivity greatly affect sampling algorithms' performance. We will not however focus on the topic in this work.

A "good enough" sample

When crawling a ground truth with a goal of obtaining a close enough sample, it is essential to have an idea of when to stop sampling. How small can a sample be? When is the sample close enough to the original graph, and what is the decisive factor of the closeness? How to measure goodness of a sample? How to measure success? These are the issues addressed by the authors of [Leskovec and Faloutsos, 2006]. They have used Kolmogorov-Smirnov statistics to compare the metrics measured. In their work they have discovered that a sample of down to 50% of the original size, kept the ground truth properties well.

The implementation we borrowed from [Kurant, 2010] follows these guidelines, which is why in our experiments we only look at the sampled graphs with minimum size of 50% of the original graph, as the algorithm implementation simply does not allow scaling any lower.

Graph access

Another issue to consider is whether or not it is possible to access the whole graph. This will highly affect the selection of the algorithm, since some of them require knowledge to all of the nodes and links in ground truth.

2.3.7 Conclusion of Graph Crawling State of the Art

Crawling is used in this thesis as a technique to sample both our ground truths. There are many different crawling algorithms. We have selected three of the most common ones - BFS, RW and MHRW to compare how well they will collect the original graphs in regards to metrics we have chosen.

As we have seen, there are many challenges with using crawling. Another technique for creating graphs we would like to look at is modeling. We will use both of the techniques to produce graphs of different sizes to measure and compare how metrics change with scale for each technique.

2.4 Synthetic graph modeling

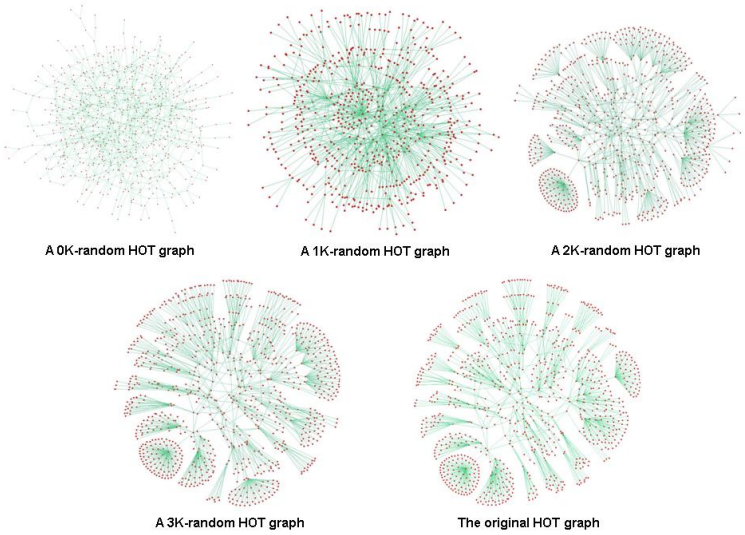
Modelling synthetic graphs is one of the approaches of social-graph analysis described by [Lee et al., 2006]. The graphs are modeled based on real graph data, modified to comply with selected features (measured by metrics) observed in real networks, for example small-world effect and the power-law degree distribution. Some of the examples of work using graph mod-

elling are the dK-series by [Mahadevan et al., 2006b], Small World networks by [Watts and Strogatz, 1998] and Nearest Neighbor model modification by [Sala et al., 2010].

2.4.1 DK-series

Figure 2.3 shows an example of graph modeling through dK-metrics, where a new model is generated for four different values of d (section 2.2 provides a more in-depth description of the dK-metrics).

Figure 2.3 Illustration of four graph models of the same original "HOT" graph shows the increasing precision of the modelled graph with the increased value for d. (Borrowed from [Mahadevan et al., 2006b])



2.4.2 Watts-Strogatz Small World Model

[Watts and Strogatz, 1998] describe in their article highly clustered systems with short path lengs they call "small-world" networks. They rewire each edge at random (using a certain probability p) for regular networks to achieve higher amount of disorder within the graph, where regular graphs have p=0 and disordered graphs have p=1. The so-called Small World networks lie somewhere in between the two poles.

2.4.3 Nearest Neighbor model

Another important study on the topic is Measurement-Calibrated Graph Models for Social Network Experiments by [Sala et al., 2010]. It discusses six graph models within three categories - "structure-driven" with focus on

structure statistics (Kronecker graphs, dK-graphs), "intent-driven" with creation process in the middle (Random Walk, Nearest Neighbor) and "feature-driven" with focus on feature statistics (Barabasi-Albert model, Forest Fire). These models are used for generating multiple synthetic graphs that substitute real graphs, comparing the results to actual Facebook data collected in 2008. The authors conclude that Nearest Neighbor graph model is the most consistent and accurate, and we have therefore initially chosen this model as the second model for our experiments. However, due to the time limit we were not able to perform the comparison using both models.

2.4.4 Conclusion of Graph Modeling State of the Art

In this section we have chosen to describe three popular models. Small-World is a well documented and tested model that is refered to by many researchers, and we have therefore chosen this model as one of our ground truths. Dk-models are usefull for this work in regards to metrics they control - we measure and compare 2K metrics for all our graphs. Nearest Neighbor was chosen as a second candidate for our ground truth, as it has previously showed to be very accurate in imitating real social graph.

By using modeling technique as a mean to create graphs of different sizes we achieve "clean" graphs without bias which is usually added by sampling algorithms. It is therefore possible to study how metrics behave on completely unbiased graphs, and then compare to metrics behavior on biased sampled graphs. We also use modeling to create one of our ground truths that we later run crawling algorithms on to see how the algorithms behave when sampling a clean graph.

There is up to date no overview over which metrics that are affected by different models. This is a definitive limitation when attempting to immitate a graph with several given properties, as is the case for our thesis' problem.

2.5 Crawling and modeling combined

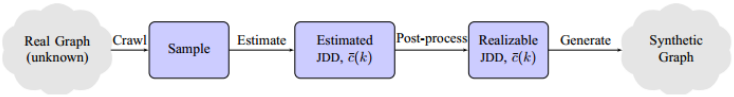
Some crawling algorithms are known to produce bias (see Section 2.3.6). Additional methods are required to correct the bias in sample produced by a crawling algorithm. One solution is to combine the two aforementioned methods into a hybrid solution where modelling is used to modify the crawled graphs.

[Gjoka et al., 2012] present a methodology to model 2.5K graphs by combining the two techniques of network sampling (crawling) and topology generation (graph modeling). The metrics chosen for comparison to the original graph are joint degree distribution (maintained by 2K-graph-models of [Mahadevan et al., 2006b]) and degree-dependent average clustering coefficient. Nodes are collected by independence sampling and random walks, and then used for generation of synthetic graphs. The results show that the

generated graphs are similar to the original for other metrics as well, such as edgewise shared partner distribution, shortest path distribution, maximal clique distribution, cycles distribution, spectrum and closeness centrality.

The process of "crawling, then modeling" approach from [Gjoka et al., 2012] is illustrated in Figure 2.4. Further details are outside of the scope of this thesis, but can be found in paper by [Gjoka et al., 2012].

Figure 2.4 2.5K crawling and modeling approach. The original graph is first crawled to obtain a smaller sample. The values for metrics (in this case Joint Degree Distribution) are then estimated, and the sample modified accordingly to produce a generated synthetic graph. Image borrowed from [Gjoka et al., 2012]



2.6 Summary of Background Chapter

In this Chapter we have looked at what social networks and graphs are, and described one example of a popular social network - Facebook. A problem with social graphs is that the graphs have become too large to experiment on, and this is why more research on proper scaling of social graphs is needed.

We found that there are mainly three different ways to achieve a smaller sample of the original graph - namely crawling, modeling, and the combination of the two. Our contribution is to compare the two most known techniques in regards to how they affect different metrics when being scaled down. There are two metrics that stand out in frequency they are measured in related work - Node Degree and Clustering Coefficient. Other popular metrics were joint degree distribution, betweenness and closeness.

This Chapter also described the challenges researchers face when crawling a graph. We use the knowledge of algorithm's bias in Chapter 4 when we interpret results to explain metrics' behaviour behaviour for different algorithms, for instance a very low clustering coefficient for BFS samples in comparison to others. We have defined the smallest sample we obtain as half of the original graph, following the ground rules set by the sampling algorithm implementation we were using, by [Kurant, 2010]. Since we use an already sampled Facebook graph as our first ground truth, and a modeled graph as our second ground truth, we did not have the issue of getting access to the real life graph data.

Modeling technique had also some challenges, although they were not as many and as obvious as the challenges with crawling. The main issue with modeling is that there is no systematic way to determine which metrics are controlled by which model. We would therefore like to contribute by

mapping some of the most common metrics as an aid in the direction of deciding which metrics to use in a given scenario.

Chapter 3

Samples, Models, and their Analysis

This chapter first presents the evalution methods we have used to compare results for metrics, namely Kolmogorov-Smirnov test, Pearson correlation coefficient and Standard Deviation. It then introduces the Facebook graph and Watts-Strogatz model we used as our "ground truth", and the graph analysis methods applied to both samples and models, explaining implementation of metrics we have previously chosen for this work.

This chapter also presents sampling algorithms we have used. We have allowed some tests to be a part of this chapter in order to justify our choice of parameters for Watts-Strogatz model in Section 3.3.2.

3.1 Evaluation approach

Evaluation of the results will be conducted through comparison of metric distributions for degree, cc, jdd, betweenness and closeness for the models and samples of different sizes.

3.1.1 Kolmogorov-Smirnov test

Kolmogorov-Smirnov statistics were applied in this work to compare metrics distribution. K-S value is calculated based on code borrowed from [“The Scipy community”, 2013b], a Python library. We define acceptable K-S distance a value within 0.1.

3.1.2 Pearson correlation coefficient

Pearson correlation was applied to measure relationship between degree and clustering coefficient metrics.

The code was borrowed from [“The Scipy community”, 2013a], a Python library.

3.1.3 Standard deviation

The implementation for Standard Deviation was borrowed from a forum [kaan, 2009], and used "as is". The code was tested by running same sequence of numbers as in the example by [Hall, 1998], which gave the same result. The formula for Standard Deviation provided by [Hall, 1998] is $\sigma = \sqrt{\frac{\Sigma(x-M)^2}{N}}$, where x stands for individual scores, M for mean (average) and N for number of scores in group.

We first calculate average metric measurement across all nodes in a given graph, then apply Standard Deviation algorithm on the average metric values from each graph to see how much the average measurements vary between different graphs.

3.2 Sample of Facebook graph as our "Ground truth"

We based our case on data collected from Facebook. Facebook is a social graph represented with user-nodes that are connected by "social relationship"-links. Links can be both directed (subscribing to updates) and undirected (requesting friendship, messaging), however in sample provided by[Gjoka et al., 2010], all links are undirected and represent friendship only. Previous work by others have made it easier to evaluate sampling results by providing already sampled sets of nodes together with metric measurements. This is especially valuable now, as more and more OSN providers are setting restrictions that prevent automated crawling on their social graphs.

The dataset we chose for our initial experiment was a BFS sample graph with 10000 nodes sampled in turn from another BFS sample provided by [Gjoka et al., 2010]. However when we attempted to construct a similar model to our graph, we couldn't get all the metrics to be similar to our Facebook sample, and therefore assumed that since it is a BFS sample, it was biased to higher degree nodes, and perhaps did not represent a social graph well enough. We have therefore chosen to both test on the sampled bias Facebook graph from [Gjoka et al., 2010], and on a "clean" Watts-Strogatz model of same size (10000 nodes).

Facebook BFS sample is undirected, static graph converted from an edge-list into a NetworkX graph for computational convenience by read_edgelist function found in [NetworkX, 2013f].

3.3 Watts-Strogatz model as our "Ground truth", model, and baseline for comparison

As reflected by the title, Watts-Strogatz model plays several roles in this work. We have first generated a graph of size 10 000 to use as:

1. ground truth for sampling experiments

2. baseline for comparison for smaller graphs of size 5000, 2500 and 1000 nodes

Similar to Facebook graph, Watts-Strogatz model is an undirected static graph, converted from edge-list into a NetworkX graph for our experiments.

We ran our first experiments on a random graph model. Watts-Strogatz implementation was borrowed from [NetworkX, 2013d] and ran with different parameters to get a wide enough range of results.

Since the average node degree for our original ground truth sampled from Facebook is approximately 46 nodes, we used the same number as value for the first parameter - the amount of nearest neighbors each node is connected to. For the parameter of probability of rewiring each edge we initially chose 0.5, to construct a Small-World graph, since 0 probability leads to regular graphs, while 1 as probability produces random graphs, according to work of [Watts and Strogatz, 1998]. Since we aim to produce a graph with properties similar to social graphs, these parameters seemed to be suitable. The third parameter, the amount of attempts to generate a connected graph, was set to 100 as a high enough number to get a resulting connected graph. The algorithm guarantees a connected graph as a result, and since it never produced an exeption during our experiments we can safely assume that the graphs produced were indeed connected.

3.3.1 Metrics under model's control

To our knowledge, the only metrics under control of the model are degree (mentioned in [Albert and Barabási, 2002]), path length and clustering coefficient (mentioned in [Watts and Strogatz, 1998]), meaning that they would not vary for models of same size and will vary predictably for models of different sizes. We need however to cross check this assumption by running the experiments for our metrics of interest (degree and clustering coefficient), and also check how the other metrics will behave between two models of same size. For probability of rewiring with value 0, the degree will be the same for all nodes. With broadening chaos probability, the degree will vary slightly, but still remain close to the original degree, described in [Albert and Barabási, 2002]. [Wikipedia, 2013c] suggests that such degree distribution is unrealistic for social graph and remains one of the model's limitations. However since we haven't found other academic sources that support this, we will have to cross check this assumption with our experiments in Section 3.3.

3.3.2 Small World vs Random Graph Models

Even though the average degree for the graph we were attempting to imitate was 46, the range of the degrees in the real graph was between 1 and 485, while in the model it was between 31 and 63, which is not very representative of the social graph we were trying to immitate. In order to improve range of degrees, we decided to see how an increased rewiring probability parameter would affect graph metrics, and have therefore constructed a random Watts-Strogatz model (by modifying the second parameter to value 1) by deducing that a random graph would provide a broader range of degrees. It did slightly improve the range of degrees, from 29 to 68, however the improvement was not that noticeable. These findings support [Wikipedia, 2013c] indication of limitations of the graph, previously discussed in section 2.4. Table 3.1 presents the K-S distance measured for each of the smaller graphs to the original 10 000 nodes size, both for Random and Small World graphs. This comparison was performed in order to find the variation that best suits our experiments and decide which value for rewiring probability parameter to use during model construction.

Table 3.1: K-S distance between graph of size 10 000 and smaller graphs for Small World and Random graphs.

Metric	Graph size in nodes	K-S distance to 10000 model	
		for Small World graph	for Random graph
Degree	5000	0.0116	0.0134
	2500	0.0083	0.088
	1000	0.0167	0.0108
CC	5000	0.056	0.6372
	2500	0.1615	0.9667
	1000	0.4518	1.0
Closeness	5000	0.9986	0.9996
	2500	1.0	1.0
	1000	1.0	1.0
Betweenness	5000	0.8613	0.8563
	2500	0.9946	0.9941
	1000	1.0	1.0
JDD 34	5000	0.0784	0.0626
	2500	0.0843	0.0585
	1000	0.0569	0.1362
JDD 45	5000	0.0133	0.0128
	2500	0.0137	0.0093
	1000	0.0169	0.0123
JDD 60	5000	0.0845	0.0743
	2500	0.069	0.0541
	1000	0.1476	0.0429

Degree

As we can see from table 3.1, there are no big discrepancies for degree metric between the two graphs.

Clustering Coefficient

The differences between the two models are much more noticeable than with degree metric. While in Small World only the smallest models are far away from the 10 000 model, in Random graph all of the smaller models are completely missing on cc measurements for the 10 000 model. The smallest one, 1000 nodes reaches the maximum K-S distance of 1 from the biggest graph.

Closeness

Closeness had a high value for K-S distance for all smaller graphs, equally high for Random and Small World graphs.

Betweenness

There was neither any noticeable difference in Kolmogorov-Smirnov distance between Random and Small World graphs.

2K

Joint degree distribution (2K) is a complex metric, producing a large number of results, describing every connected triple of nodes in the graph. Since we couldn't include all of the trippels in this thesis, we have selected three degrees to present the joint degree distribution for, and to compare with each other. First degree, 34 was chosen as one of the lowest degrees in our random graph, 45 as the degree in the middle, and 60 as one of the highest degrees.

As we can observe from the table 3.1, all the three degrees had similar values for both random and small world graph. From that we can assume that also the rest of the 2K metrics are similar for the two graphs.

Conclusion on which model is the most suitable model for our tests

The two constructed graphs were very similar in most of the measurements, however the clustering coefficient K-S distance for random graph were extremely high for all model sizes. Even though random graph had a greater range of degrees and therefore resembled more our original ground truth, the difference was not as big as opposed to clustering coefficient where Small World achieved much more stable values. We have therefore chosen Small World (with rewiring probability of 0.5) as our model of choice to run further tests on. We have also learned that randomness of Watts-Strogatz graph affects clustering coefficient, where it gets more unpredictable for smaller samples of models with high randomness.

3.4 Graph analysis

In order to analyse both sampled and modeled graphs to investigate how these will vary in scale, we have chosen a handfull of metrics, namely degree, cc, jdd, betweenness and closeness, previously described in section 2.2, to study their distribution. These metrics will indicate how well the graph will "survive" changes in size, how close it stays to the ground truth (in case of sampling), and how well it will resemble models of greater size.

3.4.1 Metrics implementation

As a starting point we chose to measure node degree and clustering coefficient, in order to easier compare our results to the results previously achieved by other authors. After the initial experiments with those most common metrics, we will include measurements of the dK-metrics (2K, the joint degree distribution), and closeness and betweenness centrality.

The metrics were described in detail in section 2.2, while the following is a description of their implementation in our experiments.

Node Degree and Clustering Coefficient

Node degree and clustering coefficient are borrowed from NetworkX library, from [NetworkX, 2013e] and [NetworkX, 2013c] respectively.

2K

2K metric, or joint degree distribution, earlier defined in section 2.2 was calculated based on the code from [Mahadevan et al., 2006a]. Since the complete calculation of 2K is three-dimentional (file is written in format of degree1-degree2-amount of edges between the two degrees) and due to the time constraints, we couldn't check the K-S distance to the ground truth for each of the degree results. Instead, we took three different degrees for each experiment and extracted the results for the ks-comparison. The reason why we couldn't choose same degrees for all the tests was because it was difficult to find three degrees that were present in each of the graphs and samples.

Closeness and Betweenness centrality

Closeness and Betweenness centrality were computed based on NetworkX implementation further described in [NetworkX, 2013b] and [NetworkX, 2013a] respectively.

3.4.2 Metrics results validation

In order to see whether we can trust our findings, the code needed some validation. This Section presents the outcome of our result validation using [Gephi, 2013] tool and code from [Mahadevan et al., 2006a]. We first plotted our Watts-Strogatz graph model of 10000 nodes into Gephi, and compared degree, clustering coefficient, closeness and betweenness calculations produced to our results for same metrics on the same graph model. Joint degree distribution was compared by replacing call for 2K calculation in our code with call to 1K calculation, which is part of the same library and only required a change of 1 parameter in our code. The idea was to produce a metric we already have by using the same library and see if it matches our results.

Clustering coefficient and degree Clustering coefficient and degree had same output when run with Gephi tool, while closeness and betweenness had K-S distance of 1 between the two result sets.

Closeness

After looking closer at the algorithms, it turned out that the one we were using for our experiments ([NetworkX, 2013b]) divided 1 with the average distance to other nodes for each individual nodes. When we divided 1 with the result we got from Gephi, the K-S distance shrinked to 0.0032. With a closer look at our two datasets with divided results, this small incoherence turned out to be due to decimal rounding.

Betweenness

Betweenness results were different due to a slight variation of algorithms, where in [wiki.gephi.org, 2013b] betweenness centrality was calculated as the sum of all the amounts of shortest paths between two nodes (a and b) for all shortest paths that passed through a particular node (c), while in [NetworkX, 2013a] this amount of shortest paths through one node was also divided by the total amount of shortest paths between nodes a and b. We did not modify gephi algorithm as we did for closeness due to the complexity of [NetworkX, 2013a] algorithm, and assume that the code is correct as the only difference to closeness computation was a call to a different NetworkX library.

Joint degree distribution

We have used 1K (degree distribution) metric as a "control metric" to verify that the code from [Mahadevan et al., 2006a] behaves the way we expected. In order to do so, we compared 1K results to our node degree results, by checking Kolmogorov Smirnov distance to the ground truth for all our samples, both for degree metric from NetworkX library, [NetworkX, 2013e] and for 1K metric from Orbis ([Mahadevan et al., 2006a]). The results were the same, and we have therefore concluded that 2K would represent joint degree distribution as claimed by [Mahadevan et al., 2006b].

3.5 Sampling Methods

This Section describes how we have generated our RW, BFS and MHRW samples.

3.5.1 Sample generation approach

We have run the three algorithms described in section 2.3 on both Watts-Strogatz graph model and Facebook BFS sample as our ground

truths. All nodes collected are unique, and sample lengths vary from 1000, 2500, to 5000 - which is 10%, 20% and 50% of the total size of our Ground Truth, respectively. We could unfortunately not sample higher amount of nodes, due to algorithm's implementation that restricted the samples to be less in size that 50% of the original graph.

The script-languages used were Python and Bash. First we ran all algorithms ten times each, selecting the best result based on Kolmogorov-Smirnov distance to Ground Truth for degree distribution (see section 3.5.2. Ten was chosen as a high-enough number to get a representative amount to select the best run from. Each time the start node was fixed for all of the algorithms to a high-degree node with degree 302 (the highest degree in the Facebook Ground Truth). A fixed start node for all experiments was chosen so that we could easier compare the results, and a high degree node was chosen since the algorithms generally perform better when they have a well-connected starting point.

Same experiment was then repeated for sample length of 2500 and 5000 nodes, each from a high degree node and a low degree one (degree 302 and degree 1). After comparing the results from both high and low degree node, we discovered that metrics were further away from the ground truth for runs with low-degree start node than with high-degree one, however the differences were not as big as we initially thought, and we have therefore only included results from the high-degree node.

Random Walk and Metropolis-Hastings Random Walk
: The code for RW and MHRW was borrowed from [Kurant, 2010] and used "as is", with a given start node, graph size and unique node collection as parameters.

Breadth First Search
: BFS algorithm was implemented in Python based on pseudocode from [Wikipedia, 2013a], modified to also take into account the graph size we wish to sample, uniqueness of nodes, and start node parameters. The uniqueness of nodes sampled is a parameter used in [Kurant, 2010] for RW and MHRW implementation, so we have also included it in our BFS implementation.

3.5.2 Comparison metric to find best run

Degree metric was chosen to be the comparison metric and decide the best run out of ten. In practice, we measured the Kolmogorov Smirnov distance (described in section 3.1) for degree metric between each run and our ground truth, compared it to the K-S distance result from the previous run, and kept the better result and sample as the "currently best one" until a next run would prove to be better.

We realize that this is perhaps not the best way, as it may not fair to the other metrics, and a good alternative could have been getting the average measurements for each metric of all the ten runs instead. However at that point it was unclear which of the two methodologies would suit better in order to produce metrics distribution that resembled somewhat the Ground Truth.

To find out whether "the best out of ten" is representative enough for the other metrics as well, we ran the same experiment for model sampling 5 times and compared the average and standard deviation for graphs of size 5000 nodes. As we can see from Table 3.2, all the metrics are fairly similar in their average, and have a very small standard deviation. Variation for betweenness was so small that we did not include it in the table. Results for joint degree distribution are not present either, as this metric was unfortunately not suited for our experiment - we elaborate further on this in Chapter 4, Section 4.1. We tested also five different runs on model sampled graphs of 2500 nodes, which showed that metrics results from our initial experiment were representative of all five runs. We have also performed five runs of same experiment on Facebook sampled graph, which similarly to model sampling produced very close average results for all of the metrics of interest. From that we can conclude that even though the choice of favorizing best out of ten runs based on degree metric may not have been the best, the metric results were representative for five other independent experiments not only for degree metric, but for average values of cc, closeness and betweenness as well.

Table 3.2: Average for metrics measurements for each of the five experiments, and standard deviation for each metric for model sampled graph of size 5000.

Experiment	Degree		CC		Closeness	
	avg	std	avg	std	avg	std
1	24.0764		0.09789		0.32656	
2	24.0664		0.09766		0.32665	
3	23.9464	0.07362	0.09747	0.00025	0.3261	0.00027
4	23.8848		0.09713		0.32593	
5	23.9604		0.09756		0.32618	

3.6 Terminology

Since there are many graphs in this thesis, it is important to give them names as well to avoid confusion. We have decided to use the following annotation to describe a graph, $G_{origin}^{algorithm}(size)$ where G

stands for graph, size for size of the graph and origin for whether the graph is modeled or sampled from Facebook. Sampled graphs have also algorithm annotation to identify whether they are BFS, RW or MHRW sample. Some examples of this annotation are listed in Table 3.3

Table 3.3: Annotations for our generated and sampled graphs.

Facebook graph as our Ground Truth	$G_{FB}(10000)$
Facebook RW sample of Ground Truth of size 2500	$G_{FB}^{RW}(2500)$
Watts-Strogatz generated graph as our Ground Truth	$G_{WS}(10000)$
Watts-Strogatz generated graph of size 5000	$G_{WS}(5000)$
Watts-Strogatz MHRW sample of Ground Truth of size 1000	$G_{WS}^{MHRW}(1000)$

3.7 Summary of Methods chapter

In this chapter we introduced our implementation of Kolmogorov-Smirnov distance, Pearson correlation coefficient and measurements of standard deviation as our tools of metric evaluation. We then presented Facebook graph sample we used as our ground truth, and Watts-Strogatz model applied for graph construction, ground truth for graph sampling and baseline for comparison for smaller models of the same type, and why we have chosen Small World implementation of Watts-Strogatz graph over the random walk. We then described our implementation and validation of degree, clustering coefficient, betweenness, closeness and joint degree distribution metrics, and how we collected the samples using the Breadth First Search, Random Walk and Metropolits-Hasting Random Walk.

Implementation of experiments was probably the most time consuming part of the whole thesis. All the experiments are summarised in table 3.4.

Table 3.4: Overview of experiments conducted in this thesis.

Graphs	**Algorithms**	**Graph size**	**Metrics**
Watts-Strogatz model	RW	1000	Degree
Watts-Strogatz model sampling	MHRW	2500	CC
Facebook BFS sampling	BFS	5000	JDD (2K)
		10000	Betweenness
			Closeness

While this chapter explained the implementation of our expreriments, the next one will present the results.

Chapter 4

Testing and Evaluation

This chapter is split into two parts - presentation and comparison of our results. Section 4.2 present how metrics' distribution changed with scale first for Watts-Strogatz graphs $G_{WS}(10000)$, $G_{WS}(5000)$, $G_{WS}(2500)$ and $G_{WS}(1000)$, then for the "clean" samples of $G_{WS}(10000)$, and lastly for sampling of Facebook graph $G_{FB}(10000)$. Before starting with experiments however, we determine in Section 4.1 which of the metrics that were not controlled by the model, and which we therefore could not use for our measurements.

Section 4.3 consentrates on our research questions from Section 1.4. First it justifies MHRW as the best performing algorithm based on the results from Section 4.2. After establishing the algorithm of choice, we compare results produced by that algorithm for $G_{FB}(10000)$ sampling, to model sampling of graph $G_{WS}(10000)$ and lastly to modelled graphs $G_{WS}(10000)$, $G_{WS}(5000)$, $G_{WS}(2500)$ and $G_{WS}(1000)$.

Before concluding the chapter, we validate our experiments, by comparing our measurements of metrics to results produced by other tools - Gephi and Orbis.

4.1 Metrics proven to be controlled by Watts-Strogatz Model

As we mentioned in section 2.4, not all metrics that we wanted to look at were controlled by model, meaning that they showed an indication to be very different between two instances of the same model. To look more systematically at the case, we created two Small World graphs with exactly same degree (46) and rewiring probability (0.5). The difference between the two graphs for clustering coefficient was surprisingly high - K-S distance of value 1. Degree and betweenness

had an acceptable K-S distance of less than 0.1, while betweenness, and joint degree distribution for degrees 34, 45 and 60 showed to be equally unaffected by model. Table 4.1 illustrates the differences in K-S distance for the two instances of the same model.

Table 4.1: K-S distance for metrics between two graphs produced by the same Small World model with parameters 46 for degree and 0.5 for rewiring probability.

Metric	**K-S distance**
degree	0.04
cc	1
betweenness	0.06
closeness	0.7
jdd 34	0.65
jdd 45	0.5
jdd 60	0.9

Based on table 4.1 we can assume that the only two metrics controlled by the model are degree and betweenness. To make sure that this assumpsion is correct, we constructed 10 equal instances of the same model, measuring standard deviation for each degree measurement from all 10 instances. Table 4.2 shows the average standard deviation for each metric.

Table 4.2: Average Standard Deviation for metrics of interest for 10 equally constructed instances of Watts-Strogatz Small World model.

Metric	**Standard Deviation**
degree	0
cc	0.03
betweenness	0.000001
closeness	0.002
jdd 34	11.18
jdd 45	208.332
jdd 60	11.817

As we can see from this overview, degree average is always the same, which is due to models algorithm that creates the desired degree distribution, with a small percentage of chaos (0.5 in our case). Clustering coefficient showed to be a false alarm from previous observation in Table 4.1, and the standard deviation showed that such high difference in measurements was only between the two first instances of the model, while for the other eight of them the values smoothed out and proved

clustering coefficient a stable metric for this model. Betweenness was as good as zero in deviation and closeness followed after, also being a false alarm from previous observations. The three degree measurements for joint degree distribution had a very high standard deviation, showing that joint degree distribution is not a reliable metric for this type of model. Based on these results we can assume that degree, clustering coefficient, betweenness and closeness are representative metrics for Watts-Strogatz model, while joint degree distribution is unpredictable and therefore unsafe to use for our metrics behaviour comparison.

Table 4.3: Average clustering coefficient for the 10 graphs constructed the same way, numbered in the order they were generated.

Graph	**Average clustering coefficient**
1	0.0045
2	0.0961
3	0.0946
4	0.0949
5	0.0945
6	0.095
7	0.09534
8	0.0955
9	0.0942
10	0.0949

For our tests we will use second generated graph, since after looking closer at the average results for clustering coefficient in Table 4.3, the first graph differed a lot from the rest for clustering coefficient (where first graph's average was 0.004, while all the rest of the nine graphs had a more stable average of 0.9).

It would be interesting to look deeper into why the results from the first graph were so different from the rest of the graphs, however we leave that for future work.

4.2 Models, model samples and Facebook samples.

This Section presents how metrics' distribution changes with scale first for Watts-Strogatz graphs $G_{WS}(10000)$, $G_{WS}(5000)$, $G_{WS}(2500)$ and $G_{WS}(1000)$, then for samples of Watts-Strogatz graph $G_{WS}(10000)$ and last for samples of Facebook graph $G_{FB}(10000)$.

Each case has a small conclusion of which of the metrics were affected by scale. For the sampling experiments we also compared performance of the three algorithms (BFS, MHRW and RW), and concluded for each case which

one best preserved the metrics' distribution. The final overview is split into the following three categories:

1. metrics that preserve the original distribution (defined by less than 0.1 K-S distance to the ground truth or less than 0.01 pearson value difference)

2. metrics' distributions that change with scale

3. metrics for which the distributions are unaffected by scale and far away from the ground truth

4.2.1 Change of Metrics with Scale for Watts-Strogatz Models

We have measured Kolmogorov Smirnov distance for metrics' distributions between $G_{WS}(10000)$ graph and the smaller graphs $G_{WS}(5000)$, $G_{WS}(2500)$ and $G_{WS}(1000)$. K-S distance results for degree, cc, betweenness and closeness are presented in Table 4.4. Pearson correlation between degree and clustering coefficient is shown in Figure 4.1.

Table 4.4: K-S distance for metrics of interest between graphs $G_{WS}(5000)$, $G_{WS}(2500)$ and $G_{WS}(1000)$ to $G_{WS}(10000)$.

Graph	K-S distance to $G_{WS}(10000)$ for			
	degree	cc	closeness	betweenness
$G_{WS}(5000)$	0.0093	0.0663	0.9994	0.8574
$G_{WS}(2500)$	0.0093	0.154	1	0.9956
$G_{WS}(1000)$	0.0083	0.4489	1	1.0

Degree

As we can see from the table 4.4, all graphs have a very small, practically zero K-S distance from the main $G_{WS}(10000)$ graph. Our initial intuition that the K-S distance to the bigger graph should increase with smaller models did not stand in this case, as all of the results show a very stable K-S value. This result is predictable, since degree was one of the parameters when constructing the model, and was given a value of 0.5 as the probability of changing each time a new graph is constructed. The degree metric is controlled by the model, and is therefore held at a more or less stable level. If we had constructed a random graph as our model ground truth, we would most likely have gotten a much greater variation in K-S distance for degree.

Clustering coefficient

Unlike the degree metric, clustering coefficient distribution gets visibly affected by size of the graph. The smaller the graph, the bigger the Kolmogorov-Smirnov distance. The smallest graph $G_{WS}(1000)$ is halfway to the top of K-S scale, with a value of approximately 0.45, however the two other smaller graphs $G_{WS}(2500)$ and $G_{WS}(5000)$ are within the acceptable range of 0.1 K-S distance to $G_{WS}(10000)$. Here we can see the trend of increased K-S distance for decreasing graph size, which is according to what we have expected. Interestingly enough, the accepted range of K-S distance for graphs of bigger size indicate that clustering coefficient metric is in fact controlled by Watts-Strogatz Small World model, which is in accordance to our previous observation from Section 4.1.

Closeness

In comparison with clustering coefficient and degree metrics, closeness distribution has a surprisingly high K-S distance (approximately 1) for all of the smaller graphs. K-S distance of value 1 means that the two samples have nothing in common, and tells us that closeness is a scale independent metric in the sence that no matter the scale, it will still not resemble itself from one sample to another. Graph $G_{WS}(5000)$ has a slightly better K-S distance than the rest, however still very close to the max distance value. Closeness distribution was kept when generating different graphs of same size in Section 4.1, however when the size became also a parameter, closeness showed to be an unreliable metric. The results also indicate that while degree distribution is being kept, the diameter of the network is of variable size, which is only natural, since the size of the graph is shrinking for each experiment. We however did not expect such drastic impact on the average distance between nodes when we produced a graph of half the size of the original. Perhaps if we had found a way to introduced diameter as an extra parameter to the model, closeness distribution would have improved.

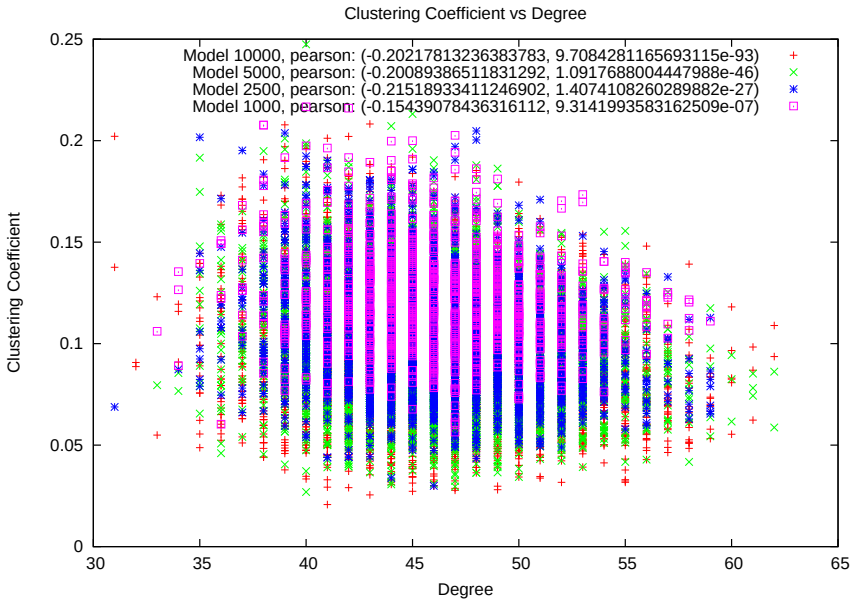
Betweenness

Similarly to average paths to each node (closeness) distribution, the distribution for amount of shortest paths (betweenness) also suffers greatly from the reduced amount of nodes in the graph. What we can see from the results in Table 4.4 is that betweenness slightly follows the assumption that K-S distance will increase with decreasing scale, however it is so high to start with (0.86 for $G_{WS}(5000)$ graph) that it also indicates that similar to closeness, it is not affected by scale.

Clustering coefficient vs Degree

Figure 4.1 shows the change of the Pearson correlation between degree and clustering coefficient metrics. The baseline correlation to compare others to is measured on $G_{WS}(10000)$ graph, with a value of approximately -0.2022. A negative Pearson correlation indicates a negative linear relation where degree increases while clustering coefficient decreases, which is according to the social graph properties previously described in Section 2.2. However, the value -0.2022 is very close to zero, meaning that the correlation is only a slight one. The Pearson correlation does not seem to get affected by size of the graph and has a value around -0.2 for all graphs, varying slightly up and down regardless of the size. From this we can conclude that Watts-Strogatz Small World model is not preserving this particular social graph property as well as we had hoped, only resembling slightly the trend of cc-degree correlation. The fact that the Pearson correlation value remains stable for most graph sizes is due to stable distribution values for degree and clustering coefficient in previously described experiments in this Section.

Figure 4.1 Change of Pearson correlation between degree and clustering coefficient metrics.



Conclusion of model generated graphs

Some of the results supported our intuition of K-S distance increasing with decreasing scale, however others indicated that the intuition might be an erroneous one. We have also learned that Watts-Strogatz Small World model does not preserve the correlation between clustering coefficient and degree, an important property of social graphs.

Table 4.5 sums up our results into the following three categories:

1. metrics' distributions that do change with scale according to our intuition

2. distributions that remain stable no matter the size of the model

3. metrics' distributions that are not affected by scale and couldn't be further away from the main model

Degree and cc-degree correlation remain stable, closeness and betweenness are the furthest they could get from the model of comparison, while clustering coefficient distribution seems to be the only one affected by scale.

Table 4.5: Overview of metrics change with scale for Small World models of different sizes.

State	**Metric**
Preserves original value	Degree CC-Degree
Changes with scale	CC
Unaffected by scale, unstable	Closeness Betweenness

4.2.2 Change of Metrics with Scale on Sampled graphs from Watts-Strogatz model

This Section presents distributions of degree, clustering coefficient, betweenness and closeness and how these change with scale by comparing samples $G_{WS}^{alg.}(5000)$, $G_{WS}^{alg.}(2500)$ and $G_{WS}^{alg.}(1000)$ to the "ground truth" graph $G_{WS}(10000)$. The "alg." annotation stands for algorithm used for sampling, which is either Random Walk, Metropolis-Hasting Random Walk or Breadth-First Search.

Table 4.6 presents the results for K-S distance to the ground truth for metrics' distributions individually, while Figure 4.2 illustrates Pearson correlation between degree and clustering coefficient.

Table 4.6: K-S distance for metrics of interest between samples $G_{WS}^{alg.}(5000)$, $G_{WS}^{alg.}(2500)$ and $G_{WS}^{alg.}(1000)$ and the ground truth graph $G_{WS}(10000)$.

Graph	K-S distance to $G_{WS}(10000)$ for			
	degree	cc	closeness	betweenness
$G_{WS}^{RW}(5000)$	0.9907	0.1312	0.998	0.8274
$G_{WS}^{MHRW}(5000)$	0.9917	0.12	0.9992	0.8308
$G_{WS}^{BFS}(5000)$	0.9406	0.2092	0.9569	0.7545
$G_{WS}^{RW}(2500)$	1	0.2322	1	0.9557
$G_{WS}^{MHRW}(2500)$	1	0.2199	1	0.9601
$G_{WS}^{BFS}(2500)$	0.9667	0.325	0.9669	0.8745
$G_{WS}^{RW}(1000)$	1	0.3389	1	0.9726
$G_{WS}^{MHRW}(1000)$	1	0.3739	1	0.9747
$G_{WS}^{BFS}(1000)$	0.97	0.414	0.95	0.8495

Degree

The K-S distance for all three sample sizes lies at approximately 0.96 for BFS and 1 for RW and MHRW. The K-S distance for degree distribution is slightly better for the bigger samples, however the difference is very marginal. Since the degree was well preserved throughout different models of various sizes, it is interesting to observe that when sampled, the degree distribution is not preserved at all. When sampled, we do not preserve the original degree of a sampled node, measuring the resulting degree based on how many neighbors of that particular node that were also sampled. BFS is therefore producing slightly better results since it is in the algorithm's nature to sample neighbors of a sampled node before going deeper into the graph. However, the difference in performance from the two other algorithms is so marginal that it is not significant for the resulting K-S distance.

So what is the reason for such difference in the degree range between the original and sampled graphs? We know that the ground truth is modelled to have an average degree of 46. It is perhaps this restriction that creates a gap in distributions, since we are collecting only a fraction of the original graph, therefore disturbing the stable average degree balance.

Clustering coefficient

Clustering coefficient K-S distance results are suprisingly low, compared to the degree results. They start with as low as 0.12 for $G_{WS}^{MHRW}(5000)$ sample, with $G_{WS}^{RW}(5000)$and $G_{WS}^{BFS}(5000)$ following close by, and slightly increase for $G_{WS}^{MHRW}(2500)$ sample to 0.21, and 0.37 for $G_{WS}^{MHRW}(1000)$ sample.

It seems that the distance gets smaller by 0.1 each time the graph size is doubled. BFS and RW show the same trend, by going from 0.21 for $G_{WS}^{BFS}(5000)$ to 0.33 for $G_{WS}^{BFS}(2500)$ to 0.41 for $G_{WS}^{BFS}(1000)$, and from 0.13

to 0.23 to 0.33 for RW samples $G_{WS}^{RW}(5000)$, $G_{WS}^{RW}(2500)$ and $G_{WS}^{RW}(1000)$ respectively.

For samples of size 5000 and 2500 the graphs $G_{WS}^{MHRW}(5000)$ and $G_{WS}^{MHRW}(2500)$ had the smallest K-S distance, while for samples of size 1000 the $G_{WS}^{RW}(1000)$ graph was better.

The low clustering coefficient results indicate that while the degree distribution was not preserved, the clustering coefficient distribution mostly remained unchanged, meaning that the neighbors of nodes were equally likely to form cliques in the sampled graphs as in the original ground truth. From this we can deduce that the amount of links between neighboring nodes stayed proportional.

Closeness

K-S distance for closeness distribution is very high (1 or almost 1) for all sizes, and there seems to be no correlation between sample size and improvement of measurements. BFS is in the lead, however the difference to the other two is minimal (approximately 0.04 difference in K-S value).

Closeness distribution was not kept when we modeled graphs of different sizes in Section 4.2.1, and has showed not to be preserved when sampling either. While for the modeled graph we explained this with the changing diameter of the graphs, in the case of sampling this assumpsion would not hold, as three different algorithms are not likely to produce graphs of so different shapes from the ground truth. This is perhaps also due to the stable average degree of 46 for the ground truth, which produces a stable amount of average distance between the nodes. When sampling, the degree distribution is not preserved, meaning that we achieve a much greater range of degrees, and the average distances from each node to all other nodes are hence modified.

Betweenness

Similarly to closeness, the K-S distance is very high for all the samples, but there is also a slight increase of K-S distance for smaller graphs. BFS starts with value 0.75 for $G_{WS}^{BFS}(5000)$, increases to 0.87 for $G_{WS}^{BFS}(2500)$ but then decreases again for $G_{WS}^{BFS}(000)$ sample to 0.85. Random walk is next best, starting with 0.827 for $G_{WS}^{RW}(5000)$ sample, increasing to 0.96 for $G_{WS}^{RW}(2500)$ sample and to 0.97 for $G_{WS}^{RW}(1000)$ sample. MHRW follows RW closely and has an increase of K-S distance from 0.831 to 0.96 for $G_{WS}^{MHRW}(2500)$ and last to 0.97.

As for the closeness and degree distributions, we can assume that this distance is due to the modification of the average degree in the sampled graphs.

Clustering coefficient vs Degree

Figure 4.2 shows Pearson correlation between clustering coefficient and degree metrics. Pearson value for the ground truth is negative, and at the same time close to zero, which means a slight negative correlation between cc and degree, where with increasing degree, cc decreases. A value close to zero means that there is practically no correlation at all. All Pearson values for samples are positive, meaning that with increasing degree clustering coefficient also increases. MHRW starts with the lowest (and in other words closest to ground truth) value: 0.05 for sample $G_{WS}^{MHRW}(5000)$ (figure 4.2a), slightly increases to 0.13 for sample of $G_{WS}^{MHRW}(2500)$ (figure 4.2b) and ends up with 0.14 for the smallest sample $G_{WS}^{MHRW}(1000)$ (figure 4.2c). Random walk is the next best, and starts with 0.06 for $G_{WS}^{RW}(5000)$, follows MHRW close with 0.14 for $G_{WS}^{RW}(2500)$ sample and 0.15 for $G_{WS}^{RW}(1000)$ sample. BFS has the furthest Pearson value from the ground truth, with 0.27 for $G_{WS}^{BFS}(5000)$, 0.31 for $G_{WS}^{BFS}(2500)$ and 0.33 for $G_{WS}^{BFS}(1000)$.

In all the three algorithms we can observe a slight increase of distance to the original value each time sample gets smaller.

As observed earlier in this Section, the degree distribution was not preserved for the sampled graphs, while clustering coefficient distribution was more stable. This explains the big gap in the Pearson values between the sampled graphs and the Ground Truth.

Conclusion of model sampling

In this Section we have learned that the degree parameter for modeled graphs is most likely affecting the behavior of distributions for degree, closeness and betweenness when a modelled graph is being sampled. We have in our experiments with sampled graphs focused on two questions :

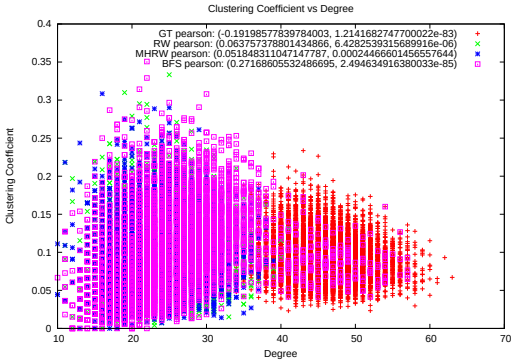
1. which metrics are affected by scale

2. which algorithm performs better than the others

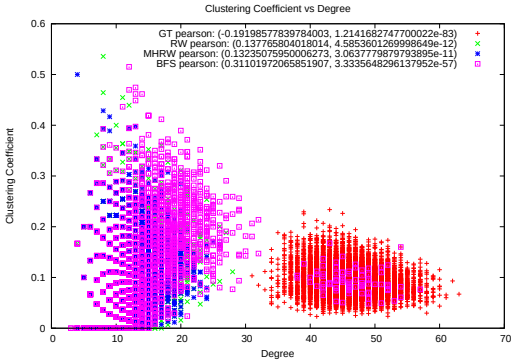
Comparison of algorithm performance:

As we can see from the results, BFS, RW and MHRW each were very close to being the best performing algorithm. BFS performed slightly better for degree, however still very far from the ground truth, and was best for both closeness and betweenness metrics. MHRW was better for clustering coefficient for higher sample sizes (both for $G_{WS}^{MHRW}(5000)$ $G_{WS}^{MHRW}(2500)$), while RW was following close by (with 0.01 K-S distance difference from MHRW for sample $G_{WS}^{RW}(5000)$, and 0.02 for sample $G_{WS}^{RW}(2500)$), even outperforming MHRW with 0.04 K-S distance

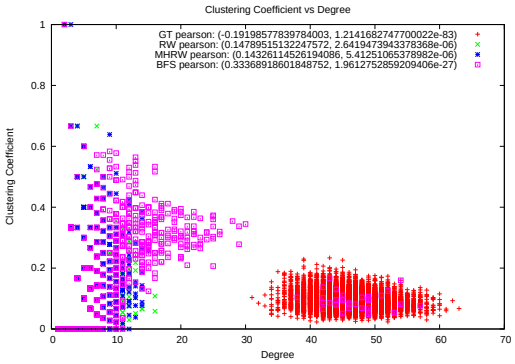
Figure 4.2 CC vs Degree comparison for Watts-Strogatz samples of size 1000, 2500 and 5000.



(a) $G_{WS}^{RW}(5000)$, $G_{WS}^{MHRW}(5000)$ and $G_{WS}^{BFS}(5000)$ samples of $G_{WS}(10000)$ ground truth.



(b) $G_{WS}^{RW}(2500)$, $G_{WS}^{MHRW}(2500)$ and $G_{WS}^{BFS}(2500)$ samples of $G_{WS}(10000)$ ground truth.



(c) $G_{WS}^{RW}(1000)$, $G_{WS}^{MHRW}(1000)$ and $G_{WS}^{BFS}(1000)$ samples of $G_{WS}(10000)$ ground truth.

difference for the smallest sample $G_{WS}^{RW}(1000)$. MHRW was the best to preserve the relation between degree and clustering coefficient. Algorithm comparison is summed up in table 4.7

Table 4.7: Overview of algorithm performance for different metrics measured on Small World samples of different sizes.

Metric	**Best algorithm**
Degree	BFS
CC	MHRW/RW
Closeness	BFS
Betweenness	BFS
Degree vs CC	MHRW

Metrics affected by scale

Table 4.8 sums up which metrics were affected by scale, which remained stable, and which showed to be unstable and unaffected by scale.

Table 4.8: Overview of metrics change with scale for Small World samples of different sizes.

State	**Metric**
Preserves original value	-
Changes with scale	CC CC - Degree Betweenness
Unaffected by scale, unstable	Degree Closeness

4.2.3 Change of Metrics with Scale on Samples of Facebook BFS Graph Sample

This Section presents our metrics measurements for degree, clustering coefficient, betweenness and closeness metrics and how these change with scale by comparing them for samples $G_{FB}^{alg.}(5000)$, $G_{FB}^{alg.}(2500)$ and $G_{FB}^{alg.}(1000)$ to the "ground truth" $G_{FB}(10000)$. "Alg." annotation stands for the sampling algorithms used, which is either Random Walk, Metropolis-Hasting Random Walk and Breadth-First Search.

Table 4.9 presents the results for K-S distance to the ground truth for metrics individually, while figure 4.3 illustrates Pearson correlation between degree and clustering coefficient.

Table 4.9: K-S distance for metrics of interest between samples $G_{FB}^{alg.}(5000)$, $G_{FB}^{alg.}(2500)$ and the ground truth Facebook sample $G_{FB}(10000)$.

Graph	K-S distance to ground truth for			
	degree	cc	closeness	betweenness
$G_{FB}^{RW}(5000)$	0.1344	0.0712	0.3879	0.2166
$G_{FB}^{MHRW}(5000)$	0.044	0.045	0.0977	0.2337
$G_{FB}^{BFS}(5000)$	0.2174	0.0938	0.6597	0.1688
$G_{FB}^{RW}(2500)$	0.0487	0.0968	0.4211	0.3633
$G_{FB}^{MHRW}(2500)$	0.1395	0.0412	0.0994	0.4456
$G_{FB}^{BFS}(2500)$	0.3042	0.1182	0.8514	0.3406
$G_{FB}^{RW}(1000)$	0.1461	0.1277	0.4278	0.4278
$G_{FB}^{MHRW}(1000)$	0.3179	0.0616	0.2173	0.2173
$G_{FB}^{BFS}(1000)$	0.3092	0.3116	0.92	0.92

Degree

The K-S distance for all samples of all sizes is relatively low, lying between 0.04 and 0.3. MHRW has the lowest distance to the ground truth, with 0.04 for sample $G_{FB}^{MHRW}(5000)$. Random Walk however achieves better results for $G_{FB}^{RW}(2500)$ sample with 0.05, a self-improvement from 0.13 for $G_{FB}^{RW}(5000)$ sample. BFS is stabilised on 0.2-0.3 K-S distance for all sample sizes. There is a trend of increased distance with smaller samples, which is followed only by MHRW algorithm. RW is more unpredictable, while BFS remains more or less stable, which is according to the node degree BFS bias previously described in Section 2.3.6.

Our intuition that K-S distance would become more increasing with decreasing size of the sample did not hold in this case, however the degree metric distribution was relatively well preserved for most samples. This is perhaps why BFS algorithm performs the worst out of the three, overestimating the node degree (as described in Node Degree Bias in Section 2.3.6), on an already bias Ground Truth (which itself is a BFS sample of Facebook graph). Since we are sampling a real graph, the degree distribution of the original graph is more naturally spread than in the previous experiments on a modelled graph, where the average degree was more or less fixed (with 0.5 chances of changing whenever a new node was added, leaving the graph slightly random, however with a restricted variety of the degree - see Section 3.3.2). It has shown to be easier for sampling algorithms to preserve degree distribution for sampled Ground Truth than for the modelled one.

Clustering coefficient

Clustering coefficient results are very low, varying between 0.04 and 0.3. K-S distance is increasing for BFS with the decreasing scale of nodes. This

behaviour is accodring to Clustering Coefficient Bias Section 2.3.6 which described that BFS tends to overestimate clustering coefficient. The bias overestimation becomes bigger when there are less nodes to sample, since the algorithm has a preference towards higher degree nodes, and therefore does not represent low-degree nodes well enough. In bigger samples the difference in amount higher and lower degree nodes collected is smaller, and the results get closer to the ground truth.

All three algorithms have acceptable distance of less than 0.1 for the largest samples of 5000 node. BFS gets left behind for the smallest sample, performing better for higher degree nodes. MHRW is best out of all three, remaining very stable and below 0.1 K-S distance for all three sizes.

Closeness

While cc and degree metrics had somewhat similar values between different runs, closeness had a much wider range of K-S distance, from 0.98 and up to 0.92 as the highest. Values are slightly increasing with scale, however they stay very close to each other within different sampling algorithms, varying only with around 0.1 between K-S distances.

MHRW is the best out of the three, while BFS is completely unable to preserve closeness for small samples, and even struggles with the biggest sample $G_{FB}^{BFS}(5000)$. Closeness however is a difficult metric to preserve, as described by [Steen, 2010], since the average closeness will vary with different graph sizes, increasing value for more users in a network and decreasing with less users. Therefore it makes sence that the variations for closeness between different algorithms are marginal, while graphs of different size are almost uncomparable. This observation stands however only for BFS and RW algorithm, most likely due to the degree bias these two are known to introduce to their samples. MHRW however is a modification of RW that slightly fixes the bias, which is why the results for closeness distribution are better preserved by MHRW.

Betweenness

Curiously enough, BFS performs better for betweenness than for both degree and closeness previously measured, and even outperforms RW and MHRW for the two largest samples. Betweenness seems affected by scale only for Random Walk, while the K-S distance is varying both up and down regardless of scale both for BFS and MHRW.

Clustering coefficient vs Degree

Figure 4.3 shows Pearson correlation between clustering coefficient and degree metrics. Pearson value for the ground truth is negative, and close to

zero, similar to our previous observation from Small World sampling experiment. All samples keep the negative value, mostly stable and close to the ground truth pearson value of -0.3, both for sample size 5000 and 2500, while worstening for sample size 1000. MHRW is best for sample of $G_{FB}^{MHRW}(5000)$, while RW is better for both smaller samples $G_{FB}^{RW}(2500)$ and $G_{FB}^{RW}(1000)$. We can as well visually confirm that all three algorithms are more or less following the Ground Truth measurements with a general trend of decreasing y-axis (clustering coefficient). This behaviour is according to what we had originally expected from our experiments. proven however only to be true for the "real-graph" sampling case.

Conclusion of Facebook graph sampling

As with experiments on sampled model, we have focused here on which metrics are affected by scale, and which algorithm performs better than the others.

Comparison of algorithm performance:

Algorithm comparison is summed up in Table 4.10

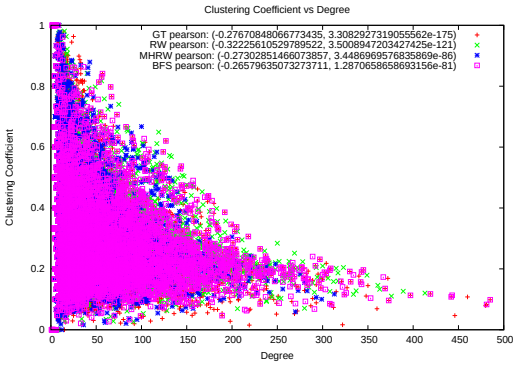
Table 4.10: Overview of algorithm performance for different metrics measured on Facebook samples.

Metric	**Best algorithm**
Degree	MHRW/RW
CC	MHRW
Closeness	MHRW
Betweenness	BFS/MHRW
Degree vs CC	MHRW/RW

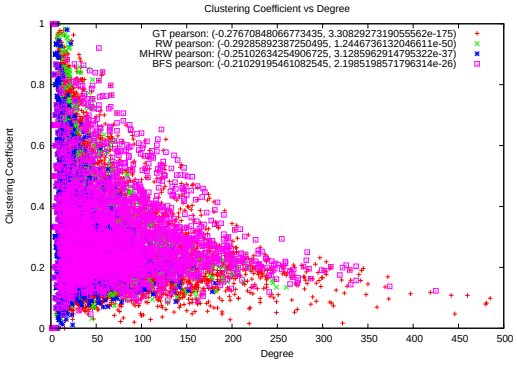
Metrics affected by scale:

Table 4.11 sums up which metrics were affected by scale, which remained stable, and which were not. For this experiment we needed to introduce a new "state" to the overview Table, in order to describe metrics that were "slightly affected" by scale. In other words, the change in K-S distance for smaller samples was present, however it was so small that perhaps for a different sample run it would have behaved differently.

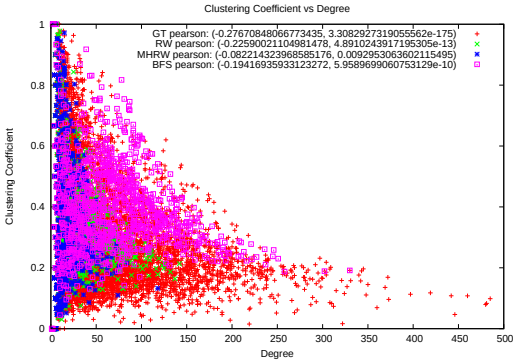
Figure 4.3 CC vs Degree comparison for Facebook samples of size 1000, 2500 and 5000.



(a) Samples $G_{FB}^{RW}(5000)$, $G_{FB}^{MHRW}(5000)$ and $G_{FB}^{BFS}(5000)$ of $G_{FB}(10000)$ ground truth.



(b) Samples $G_{FB}^{RW}(2500)$, $G_{FB}^{MHRW}(2500)$ and $G_{FB}^{BFS}(2500)$ of $G_{FB}(10000)$ ground truth.



(c) Samples $G_{FB}^{RW}(1000)$, $G_{FB}^{MHRW}(1000)$ and $G_{FB}^{BFS}(1000)$ of $G_{FB}(10000)$ ground truth.

Table 4.11: Overview of metrics change with scale for Facebook samples.

State	**Metric**
Preserves original value	CC
Changes with scale	Degree Closeness Betweenness (only for RW) CC-Degree
Unaffected by scale, unstable	Betweenness (BFS and MHRW)

4.3 Comparison of models, model samples and Facebook samples

In this section we analyze our results from Section 4 between 4 graph sizes of 10000, 5000, 2500 and 1000 nodes constructed in three different ways - by modeling ($G_{WS}(10000)$, $G_{WS}(5000)$, $G_{WS}(2500)$ and $G_{WS}(1000)$), by crawling the largest model graph $G_{WS}(10000)$ and by crawling a "real graph" $G_{FB}(10000)$ - a BFS sample of Facebook of 10 000 nodes.

In Section 4.3.1 we compare the sampled graphs based on our results from Section 4.2 and deduce which algorithm performs better in preserving different metrics with size. Section 4.3.2 presents the crawling results from the algorithm of choice, and sums up the comparison between modeling and crawling to find out which metrics were affected by scale.

4.3.1 Preferred sampling algorithm

Based on Table 4.7 and Table 4.10 from Chapter 4 we can create a new overview in table 4.12 to reflect which algorithm performs better for most metrics.

Table 4.12: Overview of algorithm performance for different metrics for both model samples and Facebook graph samples.

Algorithm	$G_{WS}samples$	$G_{FB}samples$
MHRW	CC Degree-CC	CC Degree-CC Closeness Degree
BFS	Betweenness Degree Closeness	Betweenness
RW	CC	Degree Degree-CC

As we can see from Table 4.12, BFS and MHRW are the two algorithms that outperform RW, however MHRW showed to get better results for more metrics than BFS, and this is why we have selected MHRW as our algorithm of choice.

4.3.2 Change of metrics with scale

Table 4.13 combines results from experiments conducted in Section 4.2 for metric measurements for different graph sizes both for constructed models and for sampled graphs. For sampling we have only represented MHRW results, as it has shown to outperform RW and BFS in preserving metrics measurements. A metric is counted to preserve the original value if the K-S distance between the measured graph and the baseline graph is less than 0.1, or in the case of degree-cc correlation a pearson value that differs from the original with less than 0.01.

Table 4.13: Overview of metrics variation with scale summing up the three experiments of models, sampled models and sampled Facebook graph in one.

State	G_{WS}	G_{WS}^{MHRW} samples	G_{FB}^{MHRW} samples
Preserves original value	Degree CC-Degree	-	-
Changes with scale	CC	CC CC-Degree Betweenness	CC CC-Degree Closeness Degree
Unaffected by scale	Closeness Betweenness	Closeness Degree	Betweenness

As we can see from Table 4.13, only the model G_{WS} is able to preserve the original value for degree metric, which is predictable, considering that degree was used as one of the parameters for model construction. Clustering coefficient vs degree correlation is also preserved, most likely due to non-changing degree, and a relatively small change of clustering coefficient metric for bigger models (0.07 K-S distance to the baseline for $G_{WS}(5000)$ graph and 0.15 for $G_{WS}(2500)$ graph).

Since we don't keep the original full degree of a sampled node, the degree distribution between nodes is much lower for samples than in the graph being sampled.

What is common for all the three cases is that clustering coefficient changes with scale, increasing the K-S distance to the baseline for each graph decreased in size. Clustering coefficient - degree correlation is changing with scale for both of the sampling cases ($G_{WS}(10000)$ and $G_{FB}(10000)$). In addition, betweenness changes with scale for model samples, and closeness with

degree change with scale for Facebook samples.

Models experiments have in common with model sampling that closeness is unaffected by scale and is very far away from the original baseline. Models and Facebook sampling experiments have betweenness as the metric unaffected by scale. Model samples experiment also adds degree to the list of metrics that are unaffected by scale.

4.4 Summary of Testing and Evaluation

This Chapter started with determining which of the metrics that were under control of Small World model. We have eliminated joint degree distribution from list of metrics suitable for our experiments, as it varied greatly between different instances of the same Watts-Strogatz model, and we could therefore not trust that any conclusion we based on measurements of jdd could be applied in general to Watts-Strogatz Small World graph.

We have then conducted three experiments to test how metrics vary with scale for three cases - model construction, model sampling and real graph sampling. For sampling we have compared three sampling algorithms - BFS, MHRW and RW and found out that MHRW is preserving clustering coefficient and cc-degree correlation better than the other two algorithms, while BFS is best at preserving betweenness distribution.

For the first case, the Watts-Strogatz graph models G_{WS} of different sizes, only clustering coefficient was affected by scale, while closeness and betweenness showed to be unstable and very different from the ground truth. Degree and degree-cc correlation preserved well the ground truth's original values. When sampled, Watts-Strogatz' smaller graphs ($G_{WS}(5000)$, $G_{WS}(2500)$ and $G_{WS}(1000)$) did not manage to preserve any of the original metrics measurements, however more of the metrics were affected by scale than in the models-only experiment. Degree and closeness were both unstable and unaffected by scale, while clustering coefficient, betweenness and cc-degree comparison had an increase in their K-S distance to the ground truth for each time sample size decreased. Unlike model sampling, Facebook graph sampling preserved clustering coefficient measurements of the ground truth throughout its samples. Degree, closeness, and cc-degree correlation changed slightly with scale, while betweenness showed to be unaffected by scale and with unstable K-S distance values.

For all the cases clustering coefficient distribution showed to be affected by scale, while correlation between clustering coefficient and degree was affected by scale only for the cases with samling.

Chapter 5

Conclusion

This Chapter sums up what we have done in this work in Section 5.1, answers the research questions we have asked in the beginning of the thesis in Section 5.2, concludes our main achievements in Section 5.3 as well as challenges we faced in Section 5.4 and gives some guidelines for future work in Section 5.5.

5.1 Evaluation of the thesis

The aim of this thesis was to investigate which metrics were affected by scale, either with modeling or with sampling a real graph. In this work we have performed three tests - two on Watts-Strogatz model, and one on Facebook graph. Watts-Strogatz models $G_{WS}(10000)$, $G_{WS}(5000)$, $G_{WS}(2500)$ and $G_{WS}(1000)$ were used as the "ideal case" for how metrics would behave on a "clean" graph that we could control.

$G_{FB}(10000)$ graph that we had was on the other hand a sample itself by Breadth First Search algorithm, which made it bias to higher degree nodes from the original Facebook graph, and therefore an "unclean" sample. This fact made it also interesting to test for how different sampling algorithms would behave on this sample contra the "clean" model. We also performed an experiment where we sampled $G_{WS}(10000)$ graph to see again how metrics would differ from the "ideal case" - the constructed models G_{WS}^{MHRW}, and also from the "unclean" graph sample.

With the results from these experiments at hand, we can now answer our research questions.

5.2 Anwering Research Questions

In this section we sum up our findings from Chapter 4 to answer our two main research questions:

1. "How do graph metrics change with scale?"

Out of the metrics we have measured, clustering coefficient was the only one changing with scale for all three cases. Clustering coefficient - degree correlation was also affected by scale for sampled graphs. For G_{WS} graphs where degree was fixed, the clustering coefficient vs degree correlation was preserved. Degree, betweenness and closeness results were unconclusive.

2. "Which sampling algorithm will perform better in regards to preserving metrics when scaling?"
   Metropolits-Hasting Random Walk showed to produce smaller K-S distance to the baselines of comparison for each experiment, and we therefore conclude that for preserving clustering coefficient and clustering coefficient-degree correlation (which is an important property of social graphs, previously described in Section 2.2), MHRW is the better algorithm.

5.3 Main achievements

1. We discovered that clustering coefficient distribution was affected by scale for G_{WS} graphs and G_{WS}^{MHRW} sampling cases. For G_{WS} graphs the distribution had a relatively small (less than 0.1) increase in K-S distance from $G_{WS}(5000)$ to $G_{WS}(2500)$, while from $G_{WS}(2500)$ to $G_{WS}(1000)$ the change was whole 0.3.

   K-S distance for G_{WS}^{MHRW} samples was more stable, increasing with approximately 0.1 each time the graph scale decreased with half of the nodes. Sampling real graph $G_{FB}(10000)$ showed to be least affected by scale for cc distribution, but also most stable with K-S distance changing less than 0.1 between different samples.

2. An interesting finding was that joint degree distribution was not controlled by Watts-Strogatz model at all, in contrast to degree, cc, betweenness and closeness.

3. Randomness of Watts-Strogatz graph affects clustering coefficient metric, where clustering coefficient distribution is more unpredictable for the smaller models with equal value of randomness.

4. Degree parameter of Watts-Strogatz graph affects the distributions of degree, closeness and betweenness metrics when the modelled graph is being sampled.

5. Even though Small World model has proven to be a good model for social graph construction, it failed to preserve one of the important properties of social graphs - namely the correlation between clustering coefficient and degree.

6. MHRW preserved the original metrics distribution better than RW and BFS, especially clustering coefficient and degree - cc correlation. BFS was good at preserving betweenness.

7. Watts-Strogatz model is not always a good model to use when attempting to immitate a real sampled graph. We found it difficult to create a graph based on Watts-Strogatz model that even slightly resembled a BFS sample of Facebook, since metrics such as clustering coefficient, closeness and betweenness had a very different distribution than the ground truth.

5.4 Challenges

When constructing models to resemble our Facebook sample, we first compared only degree metric. However we quickly learned that it was not sufficient to look at one metric, since low K-S distance value for one metric does not necessarily mean low K-S distance value for another, which was the case for clustering coefficient metric that showed to be a more complex metric that did not necessarily follow the low K-S distance value the other measured metrics had. Later we learned that we couldn't construct a model that resembled our Facebook sample, since it is most likely bias to the higher degree nodes, and therefore not a very good representative of a social graph. This is why we also ran sampling experiment on the biggest model, to check where the sampling would give similar results.

5.5 Future work

The field of social graph research is very wide, and we had to limit our work due to time constrains. It would have been interesting to compare the obtained results to the results of existing crawling algorithms, such as the algorithms presented by [Gjoka et al., 2010] and [Mislove et al., 2007], as well as to models from [Sala et al., 2010], [Mahadevan et al., 2006b] and [Leskovec et al., 2010].

It would also be valuable to run the sampling algorithms on a different ground truth, both different social graphs, internet topologies, and others, as well as different types of models, such as chain graph, high click graph and more, and also include more metrics, such as likelihood, k-connectivity and more. Sampling experiments included more metrics that were changing with size, however the results were not conclusive since closeness and degree was changing with scale for Facebook graph samples, while for model samples the two couldn't be further away from the ground truth, and vice versa, betweenness for model samples was affected by scale, while it was completely unaffected and far away from the ground truth for Facebook graph samples.

In this thesis we tested how standard deviation would vary for 10 instances of the same Watts-Strogatz model, and concluded with which metrics were controlled by the model (which metrics had a low value for standard deviation). It would also be interesting to check how the standard deviation would vary for models of different sizes, and see whether or not the same metrics are still controlled by the model.

In our Watts-Strogatz experiments (described in Section 4.1), we encountered a peculiar case where the metrics variation for clustering coefficient and closeness were so high that we thought we could not use these two metrics in our experiments. After constructing 10 instances of the same model and finding standard deviation for all the metrics between 10 graphs we determined that it was only the first graph that had very different results for the two metrics, while all the other nine had similar average, and therefore produced an acceptable standard deviation. It would have been interesting to explore more why two instances of Small World graph had such high variation for some of the metrics measured, unfortunately we did not have time for that.

Bibliography

[Albert and Barabási, 2002] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

[Brain, 2013] Brain, S. (2013). Facebook statistics. `http://www.statisticbrain.com/facebook-statistics/`. visited on 2013-10-29.

[Dawkins, 2006] Dawkins, R. (2006). *The selfish gene*. Oxford university press.

[Ducruet and Rodrigue, 2013] Ducruet, D. C. and Rodrigue, D. J.-P. (2013). Graph theory: Measures and indices. `http://people.hofstra.edu/geotrans/eng/methods/ch1m3en.html`. visited on 2013-01-29.

["The Scipy community", 2013a] "The Scipy community" (2013a). Graph models for online social network analysis. `http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.Pearsonr.html`. visited on 2013-05-02.

["The Scipy community", 2013b] "The Scipy community" (2013b). scipy.stats.ks_2samp. `scipy.stats.ks_2samp`. visited on 2013-09-28.

[Facebook, 2013a] Facebook (2013a). Facebook. `https://www.facebook.com/`. visited on 2013-02-22.

[Facebook, 2013b] Facebook (2013b). Open graph concepts. `https://developers.facebook.com/docs/concepts/opengraph/`. visited on 2013-01-27.

[Gephi, 2013] Gephi (2013). Gephi. `http://gephi.org/about/`. visited on 2013-10-16.

[Gjoka et al., 2011] Gjoka, M., Butts, C., Kurant, M., and Markopoulou, A. (2011). Multigraph sampling of online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1893–1905.

[Gjoka et al., 2010] Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2010). Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proceedings of IEEE INFOCOM '10*, San Diego, CA.

[Gjoka et al., 2012] Gjoka, M., Kurant, M., and Markopoulou, A. (2012). 2.5k-graphs: from sampling to generation. *CoRR*, abs/1208.3667.

[Google, 2011] Google (2011). The 1000 most-visited sites on the web. `http://www.google.com/adplanner/static/top1000/`. visited on 2013-01-26.

[Guardian, 2013] Guardian, T. (2013). Nsa prism program taps in to user data of apple, google and others. `http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data`. visited on 2013-12-08.

[Hall, 1998] Hall, R. (1998). Standard deviation. `http://web.mst.edu/~psyworld/sdsteps.htm`. visited on 2013-10-20.

[Hirst, 2010] Hirst, T. (2010). Getting started with gephi network visualisation app my facebook network, part iii: Ego filters and simple network stats. `http://blog.ouseful.info/2010/05/10/getting-started-with-gephi-network-visualisation-app-%E2%80%93-my-facebook-network-part-iii-ego-filters-and-simple-network-stats/`. visited on 2013-10-14.

[Investopedia, 2013] Investopedia (2013). Standard deviation. `http://www.investopedia.com/terms/s/standarddeviation.asp`. visited on 2013-10-20.

[kaan, 2009] kaan (2009). Display the standard deviation of a column of numbers with awk. `http://www.commandlinefu.com/commands/view/1661/display-the-standard-deviation-of-a-column-of-numbers-with-awk`. visited on 2013-10-20.

[Kurant, 2010] Kurant, M. (2010). sampling.zip. `http://mkurant.com/publications/papers/sampling.zip`. visited on 2013-02-06.

[Lee et al., 2006] Lee, S. H., Kim, P.-J., and Jeong, H. (2006). Statistical properties of sampled networks. *Physical Review E*, 73(1):016102.

[Leskovec et al., 2010] Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. (2010). Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*, 11:985–1042.

[Leskovec and Faloutsos, 2006] Leskovec, J. and Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 631–636, New York, NY, USA. ACM.

[Mahadevan et al., 2006a] Mahadevan, P., Krioukov, D., Fall, K., and Vahdat, A. (2006a). Orbis. http://sysnet.ucsd.edu/~chubble/orbis/Orbis-0.70.tar.gz. visited on 2013-03-02.

[Mahadevan et al., 2006b] Mahadevan, P., Krioukov, D., Fall, K., and Vahdat, A. (2006b). Systematic topology analysis and generation using degree correlations. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '06, pages 135–146, New York, NY, USA. ACM.

[MessyNessychic, 2013] MessyNessychic (2013). When food instagramming gets out of control. http://www.messynessychic.com/2013/04/12/when-food-instagramming-gets-out-of-control/. visited on 2013-12-05.

[Miller, 2010] Miller, T. (2010). What the hell is the social graph & why should i care about it? http://blog.search-mojo.com/2010/01/04/what-the-hell-is-the-social-graph-why-should-i-care-about-it/. visited on 2013-02-21.

[Mislove et al., 2007] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *In Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC'07)*.

[NetworkX, 2013a] NetworkX (2013a). betweenness_centrality. http://networkx.github.io/documentation/latest/reference/generated/networkx.algorithms.centrality.betweenness_centrality.html#networkx.algorithms.centrality.betweenness_centrality. visited on 2013-08-15.

[NetworkX, 2013b] NetworkX (2013b). closeness_centrality. http://networkx.github.io/documentation/latest/reference/generated/networkx.algorithms.centrality.closeness_centrality.html#networkx.algorithms.centrality.closeness_centrality. visited on 2013-08-16.

[NetworkX, 2013c] NetworkX (2013c). clustering. http://networkx.github.io/documentation/latest/reference/generated/networkx.algorithms.cluster.clustering.html#networkx.algorithms.cluster.clustering. visited on 2013-02-13.

[NetworkX, 2013d] NetworkX (2013d). connected_watts_strogatz_graph. http://networkx.lanl.gov/reference/generated/networkx.generators.random_graphs.connected_watts_strogatz_graph.html. visited on 2013-10-08.

[NetworkX, 2013e] NetworkX (2013e). degree. http://networkx.lanl.gov/reference/generated/networkx.Graph.degree.html. visited on 2013-02-11.

[NetworkX, 2013f] NetworkX (2013f). "read_edgelist". http://networkx.github.io/documentation/latest/reference/generated/networkx.readwrite.edgelist.read_edgelist.html#networkx.readwrite.edgelist.read_edgelist. visited on 2013-02-11.

[Passmore, 2011] Passmore, D. L. (2011). Social network analysis: Theory and applications. http://train.ed.psu.edu/WFED-543/SocNet_TheoryApp.pdf. visited on 2013-01-27.

[Risen Sources, 2011] Risen Sources, i. (2011). Networkx. http://risensources.com/2011/08/addictive-social-media/. visited on 2013-03-23.

[Sala et al., 2010] Sala, A., Cao, L., Wilson, C., Zablit, R., Zheng, H., and Zhao, B. Y. (2010). Measurement-calibrated graph models for social network experiments. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 861–870, New York, NY, USA. ACM.

[Sprott, 2013] Sprott, P. C. (2013). Random walk. http://scifun.chem.wisc.edu/WOP/RandomWalk.html. visited on 2013-04-06.

[StatSoft, 2013] StatSoft (2013). Pearson correlation. http://www.statsoft.com/textbook/statistics-glossary/p/button/p/#Pearson%20Correlation. visited on 2013-09-28.

[Steen, 2010] Steen, M. v. (2010). *Graph theory and complex networks: an introduction*. [s.n.], Lexington, KY.

[Wang et al., 2011] Wang, T., Chen, Y., Zhang, Z., Xu, T., Jin, L., Hui, P., Deng, B., and Li, X. (2011). Understanding graph sampling algorithms for social network analysis. In *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on*, pages 123–128. IEEE.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of "small-world"networks. *nature*, 393(6684):440–442.

[wiki.gephi.org, 2013a] wiki.gephi.org (2013a). bcloseness_centrality. http://wiki.gephi.org/index.php/Closeness_Centrality. visited on 2013-10-12.

[wiki.gephi.org, 2013b] wiki.gephi.org (2013b). betweenness_centrality. http://wiki.gephi.org/index.php/Betweenness_Centrality. visited on 2013-10-12.

[Wikipedia, 2013a] Wikipedia (2013a). Breadth-first search. `http://en.wikipedia.org/wiki/Breadth-first_search`. visited on 2013-02-11.

[Wikipedia, 2013b] Wikipedia (2013b). Bredth-first search. `http://en.wikipedia.org/wiki/Breadth-first_search`. visited on 2013-04-06.

[Wikipedia, 2013c] Wikipedia (2013c). Watts and strogatz model. `http://en.wikipedia.org/wiki/Watts_and_Strogatz_model`. visited on 2013-10-16.

[Ye et al., 2010] Ye, S., Lang, J., and Wu, F. (2010). Crawling online social graphs. In *Web Conference (APWEB), 2010 12th International Asia-Pacific*, pages 236–242. IEEE.

[Yoon et al., 2007] Yoon, S., Lee, S., Yook, S.-H., and Kim, Y. (2007). Statistical properties of sampled networks by random walks. *Physical Review E*, 75(4):046114.