# Metric Convergence in Social Network Sampling

Christian Doerr
TU Delft
Department of Intelligent Systems
2628CD Delft, The Netherlands
c.doerr@tudelft.nl

Norbert Blenn
TU Delft
Department of Intelligent Systems
2628CD Delft, The Netherlands
n.blenn@tudelft.nl

## ABSTRACT

While enabling new research questions and methodologies, the massive size of social media platforms also poses a significant issue for the analysis of these networks. In order to deal with this data volume, researchers typically turn to samples of these graph structures to conduct their analysis. This however raises the question about the representativeness of such limited crawls, and the amount of data necessary to come to stable predictions about the underlying systems. This paper analyzes the convergence of six commonly used topological metrics as a function of the crawling method and sample size used. We find that graph crawling methods drastically over- and underestimate network metrics, and that a non-trivial amount of data is needed to arrive at a stable estimate of the underlying network.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; J.4 [**Social and Behavioral Sciences**]

## General Terms

Data Analytics

## Keywords

Social Network Analysis, Data Quality, Metric Convergence, Breadth-First-Search BFS, Depth-First-Search DFS

## 1. INTRODUCTION

The proliferation of online social networking platforms with their ample availability of data opens a plethora of new research questions. With subscriber numbers in the millions, they are however simultaneously boon and bane: while enabling the study of human interactions, mobility and behavior prediction at a population scale, their massive size often restricts any analysis to only small samples of the overall networks.

The cause of this restriction is three-fold: First, the sheer size of networks limits the amount of data that can be efficiently brought in for analysis. For a network such as Facebook with more than 1 billion subscribers, a parallel crawl by 40 machines each requesting, parsing and analyzing two profile pages per second would take about 5 months of download time - after which a substantial part of the corpus is already outdated again. Second, as a protection against spammers, social media platforms usually limit the number of requests that can be made from a particular host, thereby making large crawls of topologies in most cases infeasible. Third, if a complete set was obtained, its size would often exceed the hardware requirements available to most researchers.

This naturally raises the question about the representativeness of network topologies encompassing only a comparatively limited sample of online social networks, and to what extent such samples can be used to analyze and draw conclusions about the general properties of the underlying systems. This paper will investigate this open issue for the case of Digg.com, a so-called social media aggregator where registered users submit, collaboratively curate and recommend media content on the Internet. Digg.com was chosen as a subject for a variety of reasons: First, the website was market leader among social media aggregators and until its redesign and subsequent liquidation one of the 120 most visited websites, thereby providing a diverse and large enough data source for a statistical analysis of representativeness of social network crawls. Second, the network is still manageable enough to repeatedly compute and analyze the stability of network metrics and the dynamics of different crawling methodologies. Third, using the four-dimensional data acquisition presented in [3], a complete trace of the entire social network topology was obtained, thus making an evaluation of crawling methodologies and convergence of common topological metrics possible in the first place.

The remainder of this paper is structured as follows: Section 2 summarizes commonly used crawling methodologies and previous findings into the bias of these methods. Section 3 presents an empirical evaluation about metric convergence using these commonly used crawling methodologies on the Digg.com social graph. Section 4 summarizes our findings.

## 2. THE ALGORITHMIC MECHANICS OF COMMON CRAWLING APPROACHES

Data collection in a network is steered by a data acquisition strategy, which can be roughly categorized into random sampling techniques (for example by randomly picking from a set of previously known node IDs [7]), stratified sam-

*HotPlanet'13,* August 16, 2013, Hong Kong, China.
Copyright 2013 ACM 978-1-4503-2177-8/13/08 ...$15.00.

pling or crawling. In practice, most research favors the latter as graph-traversal algorithms generate connected topologies from the very beginning, even if only a small subset of the graph has been obtained.

Among the class of graph-traversal algorithms, the classic breadth-first-search (BFS) and depth-first-search (DFS) are most widely adopted, as they are easy to understand and implement and comprehensively covered acrosss standard textbooks (e.g., [2]). From these fundamental algorithms, a number of derivations have been introduced for specialized applications and use cases, for example snowball sampling [6] which for each node only visits $n$ randomly-chosen, unknown neighbors, or forest fire [10] which probabilistically skips neighbors during its breadth-first-search. While these traversal algorithms visit nodes according to the specific order they were discovered in, other algorithms steer their search based on metrics computed at run time. Derivatives of random walk algorithms keep a transition matrix that is continuously updated based on the node degrees encountered in the search. One example of such algorithms is "non-backtracking random walk with re-weighting" [9] which eliminates the possibility of traversing back into the already known graph to increase search efficiency. Random traversal algorithms however have the disadvantage that coherent community structures of a graph only become visible comparatively late in a crawl, which makes these methods unsuited for example for user prediction and privacy research exploiting the commonalities of users within small-size clusters, unless one waits until the entire graph has been crawled or uses some method to steer its search. One of these strategies is for example mutual friend crawling [1], which utilizes link structure statistics to sequentially visit and remain as long as possible inside clusters in order to retrieve closed communities of users.

While graph-traversal algorithms such as BFS and DFS have been used for decades, an analysis about the representativeness of their output has only very recently begun with the emergence and analysis of large scale networks [8]. The fact that graph sampling can lead to a skewed result was already observed in sociological studies of small scale interaction graphs, as the discovered "your friends have more friends than you" corollary [4]. An in-depth study of this high-degree bias of BFS was conducted by Kurant, Markopoulou and Thiran [8], who were able to theoretically show the bias made when estimating the degree distribution. Their theory indicates that the average node degree only asymptotically approaches the actual degree after more than 40% of the network is sampled, and simulations of artificial 10,000-node networks match these predictions. This observation was further tested on samples of the Facebook graph in [5], which crawls of 81,000 users each obtained by scrapes of the social network. This however brings with it the complication that only open profiles and friendship relations are contained in the data set. As in some networks such as Hyves nearly half the profiles are marked as non-public, a focus on only visible relations could potentially introduce some other behavioral bias.

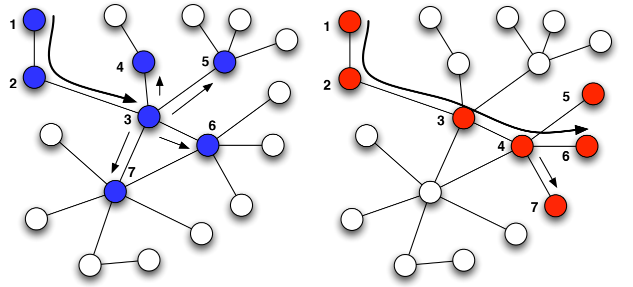
A common baseline of previous bias evaluations is the performance of a random walk on the graph, which due to its scale-free structure leads to a large estimation error of node degrees [8] due to its strong linear preference towards high degree nodes [12]. We therefore also include such random baseline, modified however with a small but important twist: our random-first-search (RFS) randomly choses a node from the list of discovered, but still unprocessed neighbors. This performs a random walk on only the discovered but not visited part of the graph without re-visiting nodes, which for a graph crawl is undesired overhead.

**Algorithm 1** BFS (DFS, RFS) Graph Traversal

```
function TRAVERSAL(G, s)
    visited ← ∅
    Q ← $List(s)$
    while |Q| > 0 do
        u ← Q.removeFirst              ▷ DFS: Q.removeLast
                                       ▷ RFS: Q.removeRandom
        visited ← visited ∪ u
        for $v \in neighbors(u)$ do
            if $v \notin Q$ and $v \notin$ visited then
                Q.addLast(v)
```

That both BFS and DFS (and as a consequence also their derivative implementations) introduce a bias is per-se not surprising if we look at the general crawling procedure depicted in Figure 1. The main conceptual difference between BFS and DFS (see Algorithm 1) can be summarized by how each algorithm chooses the next vertex to visit among those it has already discovered: As BFS processes each node in the order it was discovered, BFS will extend its search in a circular fashion from its starting point, first exploring all nodes of distance 1, before continuing to nodes at distance 2, etc. Due to the typically scale-free degree distributions of social networks, BFS will likely start at low degree node in the periphery of the graph, from where it will quickly proceed to the well-connected nodes in the center. Exploring nodes through a FIFO queue, BFS will remain in this area during the first part of the crawling process and thereby over-sample high-degree vertices. DFS on the other hand will continue its exploration process starting from the last discovered node and add newly discovered vertices to the front of its todo list. As a result, DFS will also reach the well-connected center fast but continue its search until it reaches leaf nodes in the periphery, thereby sampling the low-connected leaf nodes at a higher rate than BFS and resulting in an underestimation of nodes' degrees.

Although all previous initial work on crawling biases has focuses only on node degree, the behavior of these crawling procedures and their preference for high-degree nodes in the



(a) BFS (b) DFS

**Figure 1: Sequence of nodes processed by BFS and DFS.**

center or low-degree nodes in the periphery also heavily influences other commonly used topological graph metrics as we will show in this paper. From the schematic in Figure 1, we can already hypothesize that BFS over-estimates the graph's density while DFS will under-estimate it, and DFS exaggerates powerlaw exponents while BFS understates it. This convergence behavior of a selection of commonly used metrics will be the focus of the next section.

# 3. METRIC CONVERGENCE

This section will investigate the convergence of a number of topological graph metrics commonly used in social network analysis as a function of the amount of graph crawled. Due to size restrictions, this paper will limit the discussion to six key metrics: (1) assortativity which measures the extent pairs of nodes of similar degree connect to each other, (2) the average node degree, (3) the correlation between a node's degree and the average degree of its neighborhood, (4) the network diameter describing the network size in terms of the longest shortest path, (5) graph density and (6) a fitting of the power-law exponent on the node degrees.

The analysis of metric convergence will be conducted on the network topology of friend and follower relations within the social media aggregator Digg.com, for which a complete graph topology was collected and presented in [3]. The complete graph was obtained by the simultaneous monitoring of the Digg.com system from four different dimensions, (1) a retrieval of all content items (called "stories") submitted and listed in the platform, (2) a monitoring of all persons who commented or voted on a particular story, (3) a collection of activities performed by an individual person and (4) the tracking of all persons linked to a user by a follower or friend relation. This process was run continuously and once a new record was discovered in any of the four dimensions it was automatically added for further analysis into the other three perspectives. Since the monitoring of the website was conducted from any possible angle a registered user could have interacted with the Digg.com website, the data collection resulted in a complete view of the site's network topology. In this study we will utilize a snapshot of the Digg.com website from 2009 with about 950,000 unique users.

To analyze the convergence of above named network metrics, the Digg.com snapshot will be crawled using the previously discussed breadth-first-search (BFS), depth-first-search (DFS) and random-first-search (RFS) crawling algorithms from 100 random starting points, and tracking the development of each graph metric value. As the stability of metrics and their variance between individual runs changes over the course of the crawl, the interval at which the network metrics are assessed will also be dynamically varied. At the beginning of the crawl which is showing the highest fluctuations all graph metrics are computed every time 5000 nodes have been crawled. When the sample contains between 5% and 15% of the total network size, the metrics are computed every 15,000 nodes, and for the remainder of the traversal metrics are evaluated every 50,000 nodes. This results in an approximately equal sampling of the three regions.

From a practitioner's perspective of network crawling, it is important to note the difference between discovered and explored nodes, where the former have simply been seen by a graph traversal algorithm while the latter group has been completely processed during an iteration of the algorithm. While this seems an insignificant distinction, it implies a significant difference in practical social network crawling. When obtaining data from a social network site, a crawler typically needs to initiate individual requests for each aspect of a profile, as this keeps the processing and data transmission demands for the site's operator low and matches the vast majority of consuming application programs. For a crawler this however implies that repeated individual requests are necessary to obtain a complete view about a person's profile and friends: when requesting a user's profile page on Digg.com (but also Hyves, Twitter or Facebook), this initial response only contains the information about the single user in question. When requesting a list of friends, social networks typically only return a list of unique identifiers without providing any further context information about the friends themselves. In other words, while it is known that the user has $x$ friends there is nothing that can be said about those $x$ persons as their profile pages have to be individually retrieved. When crawling a social network, it is therefore possible to quickly compose a large list of users on a given site by retrieving the neighborhoods of only a few users (we will refer to this set as *discovered nodes*), a correct computation of most metrics in this paper will in practice however require that the profiles of the discovered users will also have been retrieved (we will refer to these as *known nodes*). We will therefore always measure the size of the crawled graph in terms of *known nodes*. Figure 2 shows the relationship between the number of crawled and thus known nodes to the number of total nodes discovered in the graph.

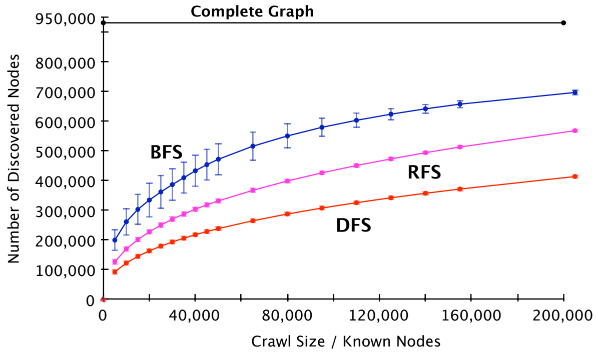


**Figure 2: Number of discovered nodes as function of crawl size for BFS, DFS and RFS.**

In the following, we will show plots depicting the development of metric values, crawled through a BFS (blue line), DFS (red line) and RFS (purple line). The line shows the arithmetic average of 100 randomly chosen starting points for a particular metric value and crawling method, the error bars show the standard deviation across these 100 runs. All plots will show the crawl size in a linear scale on the x-axis, while the metric on the y-axis will be shown either in linear or logarithmic scale to maximize readability. The solid green line displays the actual metric value for the entire Digg.com social topology if entirely crawled, the green colored area around this line displays a +/- 20% deviation from this value.

Figure 3(a)-(f) show the values and changes of assortativity, average node degree, degree correlation, density, diameter and powerlaw degree exponent respectively during the first 200,000 crawled network nodes. These initial 20% were split out separately, as most network crawls in practice are below this size and a large amount of variability takes place within this first phase, the complete view is shown in Figure 5.

As can be seen in Figure 3, DFS, RFS and BFS all converge to the final metric value, however the speed of convergence is surprisingly small and largely depends upon which metric is being used. The majority of graph traversal algorithms has not entered the +/- 20% range after the first 20% of the total network size have been crawled, there is more than a factor of 2 difference in 12 of the 18 metric/algorithm combinations after 100,000 known nodes (∼10% of the graph's size). This only improves slightly after twice the iterations, after 200,000 crawled nodes still 11 out of 18 combinations are off by a factor of 2. Despite the generally modest performance results across the three algorithms, the approximations made by random-first-search are in nearly all cases (with the exception of graph density) significantly more accurate and approaching the correct value more rapidly than those of the other two algorithms. The fluctuations in initial average metric values can be attributed to the inhomogeneity of a real-world graph in contrast to a synthetic network, where friendship creation is driven by multiple competing effects rather than a simplified generative process.

Between the in practice most commonly used BFS and DFS approaches, the results do not indicate a clear winner. For an estimation of graph density, DFS approaches the true value faster, while for an approximation of the powerlaw degree exponent BFS proves to be a superior choice. This metric-specific under- and overestimation of BFS and DFS can be intuitively explained for most of the presented metrics based on their algorithmic design as outlined in Section 2 and Figure 1. As BFS polls nodes in the order they were discovered, BFS quickly reaches the network core from its initial starting point and – having now encountered and added a long list of neighbors from these well-connected nodes – remains there for extensive periods of time. The result can for example be seen in the average node degree of the *known* nodes, which during the initial 10% of the crawl remains nearly an order of magnitude larger than its true value. This behavior has a similar impact on the graph density, which is overestimated at nearly two orders of magnitude, an estimate made on the initial data obtained from the well-connected core of the network. That BFS tends to directly reach towards the core [8] can also be inferred from its estimate of the diameter, the longest shortest path that has been found so far. BFS's diameter lifts off but than remains constant at about half its final value, again another indication that the traversal has reached from the outskirts half way across in the network into its center. The exact opposite can be associated with DFS, which builds a stack of discovered nodes exploring always the last node added first. This behavior drives this algorithm across the network core into the periphery of the graph, which can be clearly seen at the average node degree. At a value at or slightly above 2, DFS must have explored a non-trivial amount of leaf nodes in its search, if we consider that leafs have a degree of 1, the node with the next lowest possible degree (besides bridges with degree = 2) is a node connecting two leaves to the remainder of the graph already implies a degree of 3. In consequence, DFS will drastically underestimate node degree by conducting most of its search across the leaf nodes in the periphery of the graph, which can be seen in the extremely long chains of nodes being created while simultaneously also driving the estimate of the graph's density to a minimum.

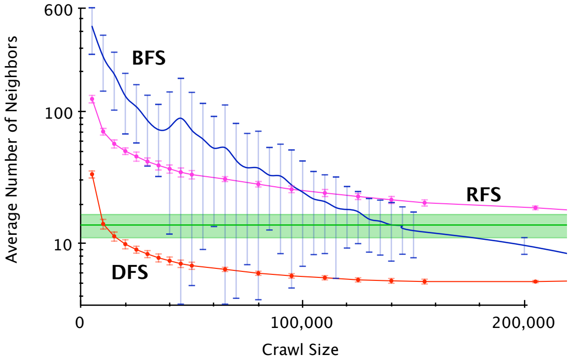
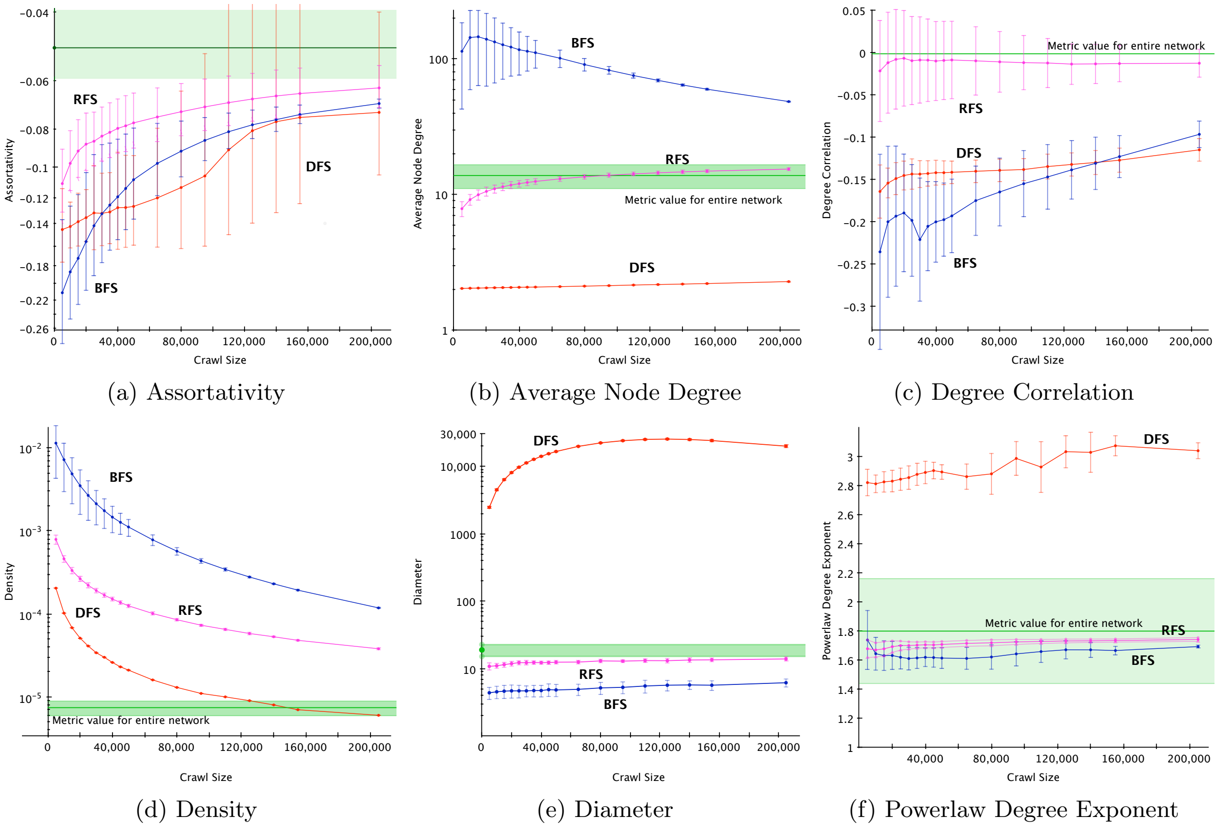


**Figure 4: The "Average Number of Neighbors" encountered in the last sampling interval - a metric to asset crawl locality.**

This behavior can be additionally visualized through the average number of neighbors (ANON), which expresses the average degree of those neighbors directly connected to the currently processed node. Figure 4 shows the average number of neighbors as a function of the DFS, BFS and RFS crawl size, binned into the current reporting interval (5,000, 15,000 or 50,000 nodes). This momentary average therefore gives a rough approximation where in the graph a traversal algorithm currently resides. High numbers of ANON indicate well-connected neighborhoods typically found within the core of a scale-free topology, while low numbers indicate a graph's periphery with a large portion of leaf nodes (where in the extreme case ANON will be close to 1). With an initial 40 fold increase over the true long-term average number of neighbors during the initial phases of the crawl, BFS clearly first explores nodes in the well connected center of the graph while only later turning into the periphery, while DFS with a ANON below half of its true value and being between 2 and 6 traverses the vicinity of the network.

While the general patterns of metric convergence can be very well explained from these behavioral characteristics of graph traversal algorithms, the extremely long convergence times and estimation performance are nevertheless remarkable, and the study was replicated using several approaches to rule out possible methodological errors. Figure 5 shows the convergence of the six target metrics for the entire crawl, and while all algorithms rebound across all metrics towards the actual final metric value, a proper convergence into the +/- 20% error margin can for some combinations take until the very end of the exploration. For many situations, the convergence is however not that grim: Both the average node degree and the powerlaw degree exponent can be properly estimated if using the right crawling strategy after 5-10% of the entire network, a solid estimate of assortativity becomes possible after approximately 30% of the entire graph. Besides the different crawling patterns, this instability is for some metrics also inherent to the metric itself, assortativity for example has been shown to be dependent on the size of a graph [11], which can be also seen in Figure 5.



(a) Assortativity (b) Average Node Degree (c) Degree Correlation

(d) Density (e) Diameter (f) Powerlaw Degree Exponent

**Figure 3: Metric convergence during initial 200,000 crawled network nodes**

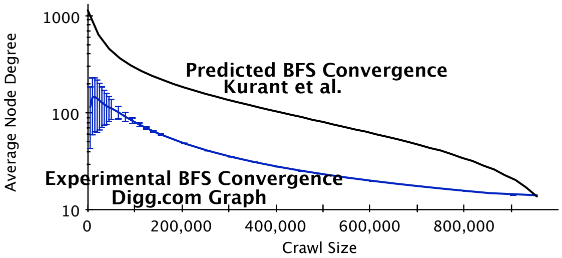


**Figure 6: Comparison of BFS convergence**

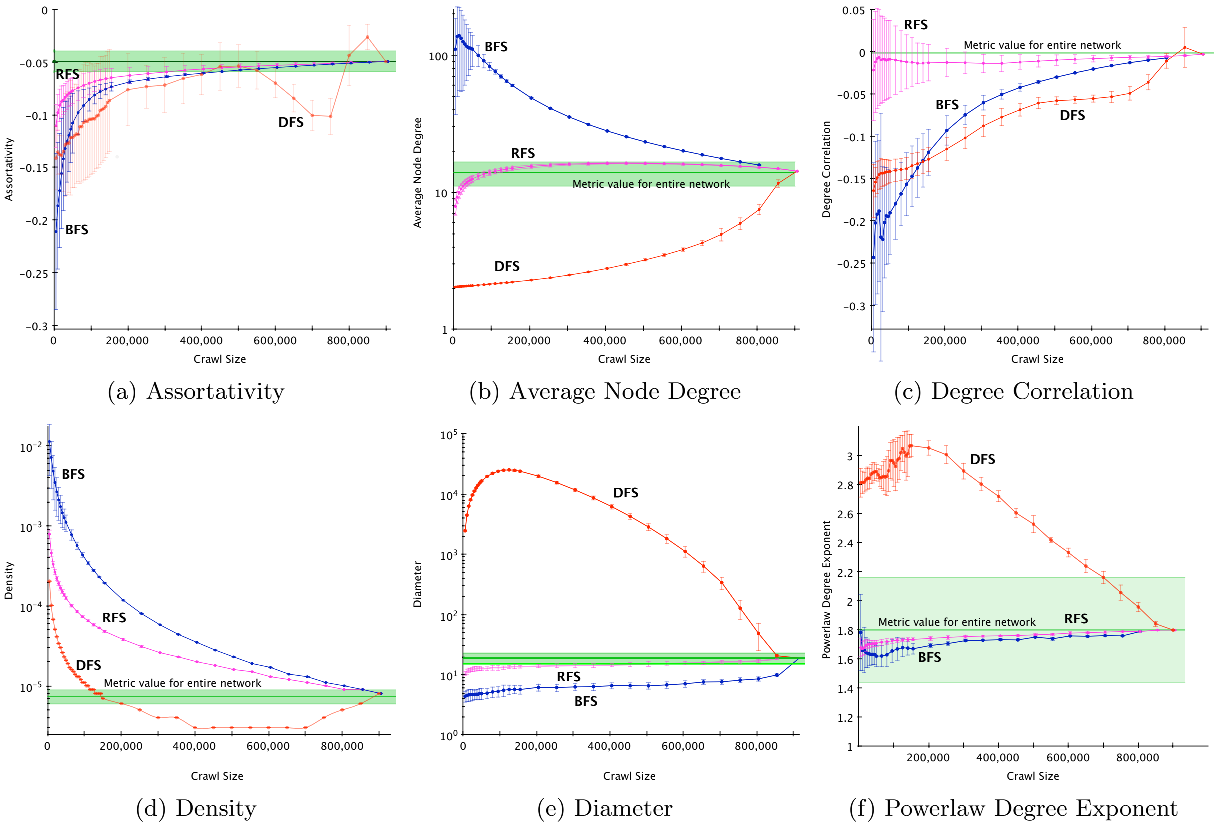
As an additional validation step, we contrasted the BFS convergence behavior and metric estimates for the average node degree as theoretically derived in [8] with those observed when crawling the Digg.com social graph. Figure 6 shows this comparison. While not providing an exact match, both approximate the true value asymptotically in a similar fashion. The empirically observed metric values are however smaller and while the cause for this deviation is not yet known at this point, we take the fact that the empirical method has a lower bias that previously predicted as a validation of the accuracy of the presented work.

## 4. CONCLUSIONS

While the availability of online social network data has enabled the investigation of new research questions, the massive size of these data sets has however also created significant logistical problems for collection and analysis. Thus, only partial crawls of these networks are used in practice.

This paper investigated the representativeness of such partial crawls for estimating commonly used topological graph metrics. We find that the convergence of these metrics varies as expected across graph traversal algorithms based on their particular crawling approaches, and that their asymptotic convergence time is surprisingly low. The results show that it is necessary in most cases to obtain more than 20-30% of the entire network to arrive at a solid estimate, for some metrics and traversal algorithm combinations nearly the entire network is necessary. This however raises the question to what extent current social network studies are able to uncover the true topological characteristics of online social networks given their limited sample sizes.

Despite this general trend, the situation is not entirely grim. If a particular use case of a data set is known a priori, it is possible to select those crawling methods that will lead to a satisfactory result even when providing a largely biased estimate for another use. If the objective is for example to assess the average node degree or powerlaw degree exponent, even a small RFS sample of 10% will provide a good foundation for this estimate. In case of a research line investigating and predicting user attributes, BFS's tendency to move in waves through the network and therefore provide a coher-



(a) Assortativity

(b) Average Node Degree

(c) Degree Correlation

(d) Density

(e) Diameter

(f) Powerlaw Degree Exponent

**Figure 5: Metric convergence during entire network crawl**

ent sample will provide a good sample to start with, even though the trace will be unsuited to make predictions about node degrees or network density. Clearly, more research will be needed to investigate this issues.

## 5. REFERENCES

[1] N. Blenn, C. Doerr, B. V. Kester, and P. V. Mieghem. Crawling and detecting community structure in online social networks using local information. In *IFIP TC 6 conference on Networking*, 2012.

[2] H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2001.

[3] C. Doerr, N. Blenn, S. Tang, and P. Van Mieghem. Are friends overrated? A study for the social news aggregator digg.com. *Computer Communications*, 35(7), 2012.

[4] S. L. Feld. Why Your Friends Have More Friends Than You Do. *American Journal of Sociology*, 96(6):1464–1477, 1991.

[5] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM*, 2010.

[6] L. A. Goodman. Snowball sampling. *The Annals of Mathematical Statistics*, 32(1), 1961.

[7] B. Krishnamurthy and C. E. Wills. Characterizing privacy in online social networks. In *WOSN*, 2008.

[8] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of BFS (Breadth First Search). In *ITC 22*, 2010.

[9] C.-H. Lee, X. Xu, and D. Y. Eun. Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling. In *SIGMETRICS*, 2012.

[10] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *KDD*, 2005.

[11] N. Litvak and R. van der Hofstad. Uncovering disassortativity in large scale-free networks. *Phys. Rev. E*, 87:022801, Feb 2013.

[12] L. Lovasz. Random walks on graphs: A survey. *Combinatorics*, 2:1–46, 1993.