Algorithms for Analyzing and Mining Real-World Graphs

Takes, F.W.

8

Mining User-Generated Path Traversal Patterns in an Information Network

This chapter studies patterns occurring in user-generated clickpaths within the online encyclopedia Wikipedia. The clickpath data originates from over seven million goal-oriented clicks gathered from the Wiki Game, an online game in which the goal is to find a path between two given random Wikipedia articles. First we propose to use node-based path traversal patterns to derive a new measure of node centrality, arguing that a node is central if it proves useful in navigating through the network. A comparison with centrality measures from literature is provided, showing that users generally "know" only a relatively small portion of the network, which they employ frequently in finding their goal, and that this set of nodes differs significantly from the set of central nodes according to various centrality measures. Next, we consider so-called frequent traversal graphs, i.e., graphs that arise from considering the nodes and edges of the top-k frequent path traversal patterns. We demonstrate how a small set of patterns is enough to obtain a subgraph with structural properties similar to that of the original graph, showing that users are able to identify a small yet efficient portion of the graph that is useful for successfully completing their navigation goals. This chapter is based on:

- F. W. Takes and W. A. Kosters. Mining user-generated path traversal patterns in an information network. In *Proceedings of the IEEE/ACM International Conference on Web Intelligence (WI 2013)*, pages 284–289, 2013

8.1 Introduction

A large part of the gigantic amount of information that is nowadays available is organized in some sort of *network* structure. Examples include the world wide web, an online social network or an information network such as Wikipedia. In these networks (or graphs), each node represents an entity or a piece of information, and each link represents a tie or relationship between two entities. An important task that human users perform on a daily basis, is *searching* for a piece of content within such a network. Although search engines can often assist the user in performing such a search task, *navigating* to the desired page by means of clicking the links between the nodes in the network is still a common activity, as sometimes search engine performance does not exactly meet the user's needs [133]. In such cases, the user will have to reach the correct page by traversing hyperlinks that exist between the pages in the network, forming a path towards the correct piece of information. Throughout this chapter we consider the task of mining *traversal patterns* that occur within these types of clickpaths. The obtained patterns are useful for understanding pathfinding strategies in networks, and may even be useful in getting a better understanding of human search behavior in general [62].

The data used in this chapter originates from the Wiki Game, an online game in which the main task is to link two given random pages on Wikipedia. Employing his perception of the structure of the network, a user has to find his way to the goal article by clicking the directed links that exist between the various articles in the Wikipedia graph, essentially generating a goal-oriented clickpath. We will consider a newer version of the Wiki Game dataset introduced in Chapter 7, containing more than one million clickpaths, comprising a total of seven million goal-oriented clicks on Wikipedia pages. It is important to note that these clicks are fundamentally different from simply counting the number of visits to a certain page, as these counts would for example also include visits that immediately reach the desired goal page, for example via a search engine. Instead, the clickpaths that we will study consist of Wikipedia pages and links between pages that were actually considered useful, by the user, in *traversing* the network.

We will use node-based traversal patterns to address a problem within the field of network analysis called *node centrality*, defined as the importance of a node within the network. So-called *centrality measures* are widely used to assess this issue of node centrality, and examples include PageRank [107] as well as centrality measures that originate from the field of social network analysis such as degree centrality, closeness centrality and betweenness centrality [26]. While the aforementioned centrality measures all employ the structure of the network to assess the importance of a node, none of them incorporates the human perception of the information incorporated in

the network. As it is ultimately the user who is going to assess whether or not a page is actually relevant, one could say that it is not the structure of the network which should serve as the basis of the centrality measure, but it should instead be the user's perception of the network that is going to determine the importance of a node. It may very well be that certain structurally central nodes in the network are not considered important or useful by the user, and vice versa. Therefore we introduce a user-defined measure of centrality based on frequently traversed nodes, arguing that a page is important if it proves useful in navigating through the network. Especially in networks where the user perception of the data plays a central role, such as in the world wide web, or in an information network, we believe that a user-defined measure makes more sense than a conventional user-insensitive approach. Furthermore, we introduce the measure of subgraph centrality which determines the centrality of a group of connected nodes with respect to the rest of the network, allowing an experimental verification of the quality in terms of ease of navigation of the user-perceived central nodes.

The rest of this chapter is organized as follows. In Section 8.2 we discuss some definitions and introduce our dataset. After discussing related work in Section 8.3, we introduce node-based patterns and our user-defined measure of centrality in Section 8.4. In Section 8.5 we analyze different types of subgraphs derived from frequent traversal patterns, and we perform experiments demonstrating the successful use of these subgraphs by humans. Section 8.6 concludes the chapter and provides suggestions for future work.

8.2 Preliminaries

This section starts with some basic definitions regarding graphs and paths that will later on allow us to precisely define our path traversal patterns and various derived measures. We also describe the clickpath dataset to which we will later on apply our path traversal pattern mining techniques.

8.2.1 Wikipedia graph

We will model the information network Wikipedia as a directed graph $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ directed links between pairs of nodes. The indegree $indeg(v)$ of a node $v \in V$ is equal to the number of incoming links of v, and similarly $outdeg(v)$ denotes the number of outgoing links. We define a *path* as a vector p of visited nodes, where for each subsequent node pair $(v_i, v_{i+1}) \in p$ there exists a link $e = (v_i, v_{i+1}) \in E$ in the original graph G. The *path length* is then equal to the number of links that was traversed to get from the first to the last node in the path. We define

the *distance* $d(u, v)$ as the length of the shortest path between nodes u and v, meaning the minimum number of links that has to be traversed to get from u to v. If there is no path between u and v, then $d(u, v) = \infty$. In such cases, the graph has multiple strongly connected components, meaning that some nodes are not reachable from every other node by considering the directed links between the nodes. This does not necessarily mean that there are multiple weakly connected components, which we define as maximal sets of nodes such that every node can reach every other node by means of traversing the links between the nodes, regardless of the direction of the link. For convenience in later definitions, we denote the number of shortest paths between two nodes by $\sigma(u, v)$, and the number of shortest paths from u to v that runs through node w as $\sigma_w(u, v)$.

In this research, we use a Wikipedia graph consisting of the pagelinks from the English version of DBpedia version 3.7 and 3.8 (see [10] or `http://dbpedia.org`), which were mined from the original Wikipedia datasets in 2011 and 2012. We mention that by only considering actual pagelinks and ignoring links to special pages or external websites, each page represents an actual piece of information within the information network. Although the used Wikipedia graph is a bit newer than the version presented in Table 7.1 in Chapter 7, some more pruning of "special" Wikipedia pages was done, resulting in a graph with $n = 3,416,126$ nodes and $m = 83,271,539$ directed links, and further statistics similar to what we presented in the previous chapter.

8.2.2 The Wiki Game dataset

The clickpath data used in this chapter is based on clicks made by users of the Wiki Game (`http://www.thewikigame.com`). In this game, users are assigned the task of connecting two given random articles on Wikipedia by traversing the links that exist between Wikipedia articles. For additional information and an example of this game, the reader is referred to Section 7.2.3 of Chapter 7. Compared to the previous chapter, we study a newer version of the Wiki Game dataset. Furthermore, the focus is on the actual completed clickpaths, and failed paths are ignored.

Property	**Value**
All user-generated paths	$3,219,641$
Clicks in all user-generated paths	$17,151,824$
Percentage successful	35.3%
Successful paths	$1,137,337$
Clicks in successful paths	$7,135,060$

Table 8.1: The Wiki Game dataset used in Chapter 8.

The dataset used in this chapter consists of clickpaths generated between 2009 and 2012, where one clickpath corresponds to a played game (or task), which is is essentially a (start, goal) pair between which a path has been formed. In this chapter, we will only consider paths with a length between 3 and 20, thus filtering out non-serious attempts. A total of $3,219,641$ paths was generated, consisting of $17,151,824$ clicks in total. Of these tasks, little over one third was successfully completed, which is the part of the dataset that we consider in this chapter. This results in a dataset of $1,137,337$ clickpaths consisting of a total of $7,135,060$ clicks. The statistics discussed above are summed up in Table 8.1. Figure 8.1 shows the relative frequency of the lengths of all user-generated paths, as well as that of the computed shortest path lengths of the tasks.

8.3 Related work

Path traversal patterns in a hyperlinked environment have been a popular subject of study since the introduction of the web [30]. A lot of work has been done on mining the top-k frequent traversal patterns [93], often by using algorithms from the field of frequent itemset mining [53]. With the enormous amount of web traffic taking place these days, studying path traversal patterns from a stream has also become a useful

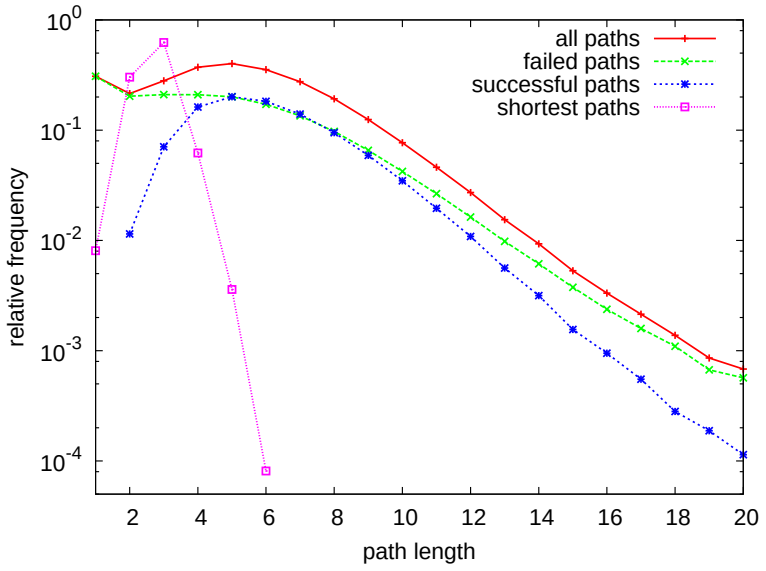

Figure 8.1: Relative frequency (vertical axis, logarithmic) of various path lengths (horizontal axis) of the filtered dataset.

task [92]. Most of the research in which clickpaths are analyzed within a confined environment considers weblogs from a particular website [1]. This chapter differs from such studies in a sense that all clicks in our dataset are goal-oriented and clicks are identifiable as one unique topic, namely the subject of the Wikipedia page. An overview of additional related work, for example on analysis of Wikipedia itself, can be found in Section 7.3.

In Chapter 7, we have investigated the difficulty of forming a path between two given random pages, showing that in Wiki Game, the indegree of the goal page as well as the reversed neighborhood, both local properties of the goal page, are good predictors of the difficulty of performing such a path traversal task. We have also demonstrated how the start page is of little influence as the user just navigates away from it quickly in search for a hub. Whereas the previous chapter only considered path traversal success or failure, in this chapter, we consider the patterns that arise from the actual clicks made by the users.

8.4 Path traversal patterns

In this section we will first introduce three types of path traversal patterns, after which we look in detail at node-based traversal patterns, and how these patterns can serve as a basis of a user-defined measure of centrality. We will compare this new measure with centrality measures from literature, that we briefly describe in Section 8.4.2.

8.4.1 Patterns

Given a dataset P consisting of a large number of clickpaths, we are interested in *patterns*, i.e., observable phenomena that occur more frequently than expected. For our clickpath dataset, it is possible to distinguish between the following frequencies in order to define our patterns:

- *Node traversal frequency*: the number of times a node v occurs in all paths p from P.
- *Edge traversal frequency*: the number of times an ordered pair of subsequent nodes (v_1, v_2) occurs in all paths p from P.
- *Subpath traversal frequency*: the number of times an ordered sequence of three or more subsequent nodes $(v_1, v_2, v_3, \ldots)$ appears in all paths p from P.

Obviously, relaxing the definition of subpath traversal frequency to length two or one, yields the definitions of respectively edge and node traversal frequency. Similar to the

definitions often given in the area of frequent itemset mining [53], we call an observation *frequent* if it occurs more often than a certain threshold $\theta > 0$ amongst all paths, allowing the definition of our patterns: *frequent nodes*, *frequent edges* and *frequent subpaths*. For a given threshold θ, every node within a frequent edge and every edge within a frequent subpath, is also frequent. We define the set of top-k frequent patterns as the set of $k \geq 1$ patterns with the highest frequency, allowing us to again define derived sets called *top-k frequent nodes*, *top-k frequent edges* and *top-k frequent subpaths*. The most simple patterns based on frequent nodes are further discussed in this section, whereas graphs derived from more complex traversal patterns are considered in Section 8.5.

8.4.2 Centrality measures

Node centrality as the importance of a certain node in the graph. A centrality measure M returns the centrality $C_M(v)$ of a node $v \in V$. We consider the following (existing) centrality measures, somewhat ordered by their complexity in terms of computation time:

Indegree centrality

$$C_{indeg}(v) = \frac{indeg(v)}{n-1}$$

Closeness centrality

$$C_c(v) = \frac{1}{\frac{1}{n-1}\sum_{w \in V} d(v,w)}$$

PageRank

$$C_{PR}(v) = PR(v)$$

HITS

$$C_{HITS}(v) = a(v)$$

Betweenness centrality

$$C_b(u) = \sum_{\substack{v,w \in V \\ v \neq w, u \neq v, u \neq w}} \frac{\sigma_u(v,w)}{\sigma(v,w)}$$

A discussion and more elaborate definition of these measures is given in Section 5.4.1 and Section 6.3. Each of the centrality measures results in a number between 0 and 1, where a higher score indicates that the node is more central. For convenience, we normalize the centrality values such that the most central node has a centrality value of 1. Clearly, distance based measures do not perform well when there is more than

one connected component. Therefore we will only consider the largest strongly connected component of the Wikipedia graph. We believe that we have covered the most common and applicable ones in this subsection, using similar arguments regarding the type of measures as presented in Section 5.4.1.

8.4.3 User-defined node centrality

Recall from Section 8.4.1 that considering the *top-k frequent nodes* means that if we sort the list of nodes by their node frequency value, we consider the k nodes with the highest frequency. For our clickpath dataset, this means that we are looking at the k nodes that were most frequently used to traverse the graph. This list is actually quite interesting, as it indicates which k nodes are considered important, by the user, in navigating through the graph. We use this data as a basis for our user-defined measure of centrality, proposing to count the number of clicks that an article v received (denoted by $clicks(v)$) and divide it by the total number of clicks made in order to obtain our user-defined measure of centrality:

User-defined centrality

$$C_{ud}(v) = \frac{clicks(v)}{\sum_{w \in V} clicks(w)}$$

To get an idea of the values returned by this function, Figure 8.2 shows the frequency of each node traversal count over all nodes in the graph. The distribution follows a clear power-law, meaning that many nodes are visited only a few times, and a few nodes are visited quite often. We are obviously interested in the tail of the distribution: the set of nodes that is visited very frequently.

8.4.4 Measure evaluation

Assessing the quality of a centrality measure is not a trivial task, and often comes down to simply comparing one centrality measure with another centrality measure. An alternative would be to have a subjective evaluation done by a human, and then determine the extent to which the ranking produced by the centrality measures resembles the user's perception of the importance of these nodes. An example of a more authoritative ground truth for centrality is provided in Chapter 6 in the context of online social networks, where celebrities have a special labeling created by the administrators of the network, reflecting the celebrity status of the real-world person behind the profile. However, often researchers rely on manual inspection of the top-k most central nodes [106], or simply compare their measure with other existing centrality measures [22]. If we are only interested in the relative ranking of entities in

two top-k lists, measures such as Kendall's tau and Spearman's weighted footrule can be used [79].

In our experiments, we will use two different ways of comparing centrality measures, as suggested in [22] (though in a somewhat different setting). Often, a centrality measure is used to find the top-k most central nodes, and we mention that the evaluation techniques that we discuss here are designed such that thus only the top-k nodes are evaluated. A rather basic technique is to compare top-k nodes of two centrality measures and determine the percentage of nodes that overlap. For example, for $k = 1$, we simply verify whether the most central node is equal for both measures. We call this measure *top-k precision*, defined as follows:

$$\text{top-}k \text{ precision} = \frac{|A_k \cap B_k|}{k}$$

Here, $A_k, B_k \subseteq V$ represent the sets of top-k nodes returned by centrality measures A and B. Alternatively, when the actual centrality value of the top-k nodes is also of importance, we can look at the correlation between the centrality values in two lists of nodes. We call this measure *top-k correlation* and define it as the Pearson correlation coefficient between the centrality values of the two methods. Important to note here is that measure A is considered as the ground truth: we compare the centrality values of the top-k nodes of measure A with that of measure B.

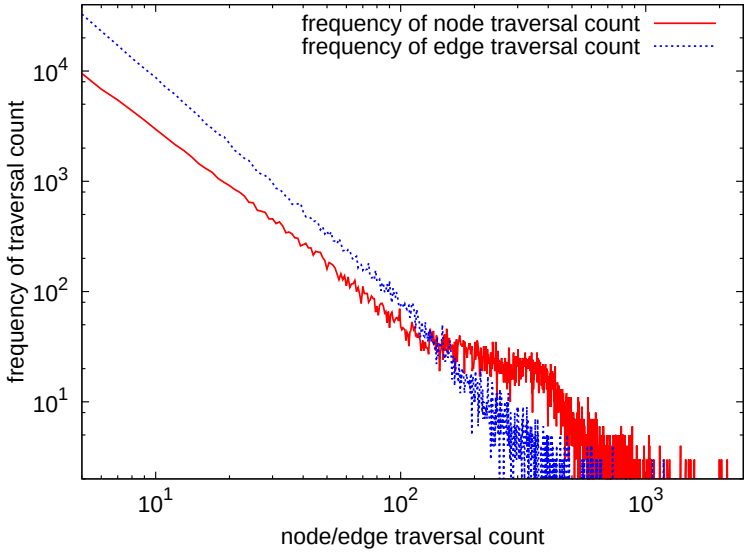

Figure 8.2: The frequency (vertical axis, logarithmic) of different node and edge traversal counts (horizontal axis, logarithmic).

8.4.5 Experiments

In this section we use the user-defined measure of node centrality introduced in Section 8.4.3 as a ground truth for comparing the centrality measures listed in Section 8.4.2. We compare the different measures up to $k = 250$, based on an evaluation using both top-k precision (see Figure 8.3 and Table 8.2) and top-k correlation (see Table 8.2).

We note that for small values of k, big deviations for the top-k precision measure can be observed, which is due to the fact that with a low value of k, one mismatch has a relatively high influence on the actual percentage. In our experiments we also found that it is important not to lose the directed aspect of the Wikipedia network, as otherwise overview pages containing listings of events or people will be ranked too high. This is also the reason why both outdegree centrality and the HITS algorithm using the hub score instead of the authority score did not produce meaningful results.

Looking at the performance of the different centrality measures in Table 8.2, we can generally conclude that PageRank gives not only the highest, but judging from Figure 8.3 also gives the most consistent results when top-k precision is considered. Indegree centrality is a good second choose if top-k correlation is important. We mention that for values greater than $k = 250$, a somewhat consistent precision is observed. Altogether, it appears that centrality measures are able to explain only roughly half

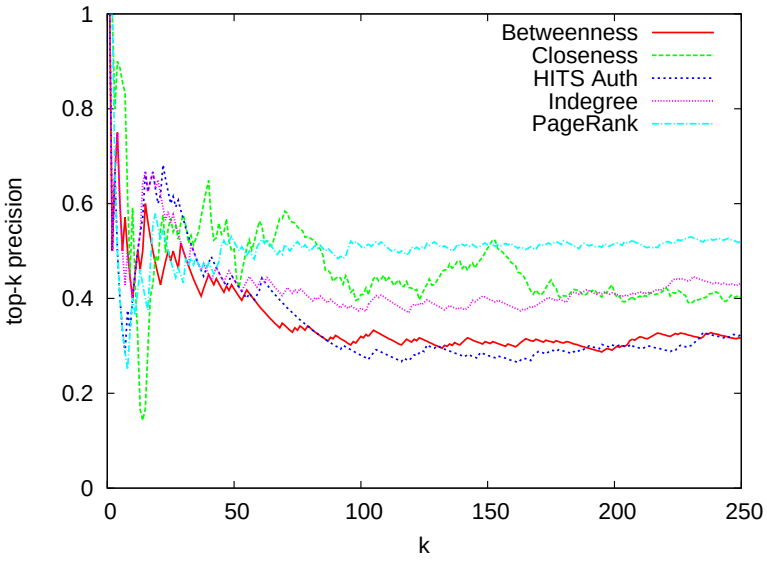

Figure 8.3: User-defined top-k precision (vertical axis) for different k (horizontal axis).

Measure	**Top-k precision**	**Top-k correlation**
User-defined	1.00	1.00
PageRank	0.51	0.76
Closeness	0.49	0.53
Indegree	0.37	0.83
Betweenness	0.32	0.71
HITS	0.28	0.62

Table 8.2: Comparison of centrality measures with user-defined centrality for $k = 100$.

of the nodes that are frequently used by humans to traverse the graph. This may lead us to believe that either humans are able to assess half of the central nodes in the graph, or that existing centrality measures are simply not able to produce the portion of nodes which is considered useful by the user. In the latter case, the only remaining question is then whether or not the set of nodes returned by the centrality measures is better or worse at ensuring that a large portion of the graph is easily reachable and thus useful for completing navigation goals. We will try to answer this question in the next section by looking at global properties of the path traversal patterns.

8.5 Global patterns

In this section we consider the global properties of the frequent patterns, creating subgraphs of the original network by considering the frequent node and edge traversal patterns.

8.5.1 Frequent traversal graphs

We define a *frequent traversal graph* as a graph consisting of frequent traversal patterns, distinguishing between two types of graphs:

- *Node-based frequent traversal graph*: the subgraph consisting of all nodes $v \in V$ and their connecting edges for which it holds that v is traversed more often than a certain threshold $\theta > 0$.
- *Edge-based frequent traversal graph*: the subgraph consisting of all links $(u, v) \in E$ and their node endpoints for which it holds that (u, v) is traversed more often than a certain threshold $\theta > 0$.

Analogously to the definitions given in Section 8.4.1, we can define top-k node-based and edge-based frequent traversal graphs, consisting of the top-k most frequently

visited nodes and edges, respectively. Iterating over increasing values of k then yields node-based and edge-based evolving graphs.

Some properties of these two types of subgraphs are shown in Figure 8.4 and Figure 8.5 for respectively the frequent nodes and frequent edges of our Wikipedia clickpath dataset. We note that when considering frequent edges, the average distance between two nodes quickly (from roughly $k = 5,000$ onwards) resembles that of the original Wikipedia graph (4.55), and then remains surprisingly stable as k increases. The node-based frequent traversal graph does not resemble the distance distribution of the original graph, as this type of subgraph also contains many edges that were not actually traversed, but are simply present between the frequent nodes in the original graph, creating many more connections between the nodes than were actually traversed. Indeed, considering only the edge-based frequent patterns might make more sense, as the user apparently "knew" these exact links, and not just the nodes. The findings presented here may indicate that the user is able to select a representative portion of the edges (and by that a portion of the nodes). In the next section, an attempt is made to measure the effectiveness of this central portion of nodes.

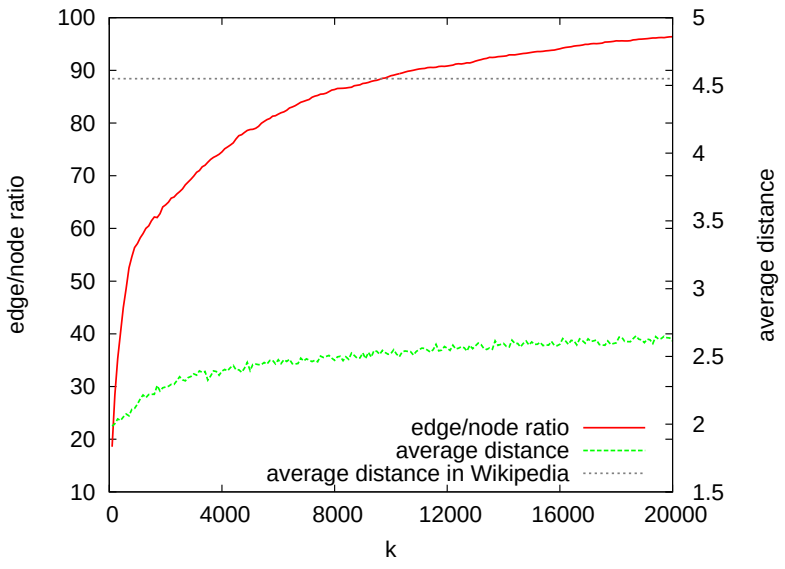

Figure 8.4: Values (vertical axes) of different properties of the node-based frequent traversal graph for different k (horizontal axis).

8.5.2 Subgraph centrality

The final question which we aim to answer in this chapter, is whether or not the frequent traversal graphs are actually better or worse than graphs derived from traditional centrality measures in terms of being able to quickly reach a large portion of the original graph, and thus ensuring ease of navigation. To do this, we introduce the measure of *subgraph centrality*, which we define as the centrality (according to some existing measure, in our case closeness centrality) of a *set* of nodes, namely the set of top-k nodes obtained through a centrality measure. To determine the centrality of this set of nodes, we merge the set of top-k frequent nodes into one node, realizing the equivalent of setting the weight of all edges between frequent nodes to zero.

In Figure 8.6 we show for increasing k the subgraph centrality values derived from the frequent nodes in the user-defined measure and the PageRank centrality measure. We have chosen to provide a comparison with PageRank and indegree because they performed best in terms of precision and correlation according to our experiments in Section 8.4.5. We observe how the subgraph centrality of the user-defined frequent traversal graph compares quite well to that of the PageRank subgraph, which indicates that the user is able to select a portion of nodes which in terms of reachability is equal to that of a centrality measure. For $k > 1,200$, the quality of the user-defined centrality is even higher than that of PageRank, suggesting that users are able to select a portion

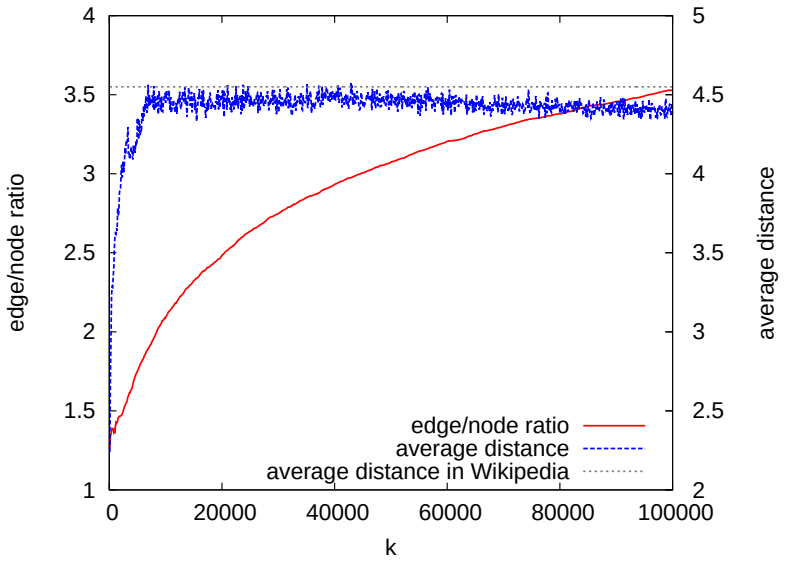

Figure 8.5: Values (vertical axes) of different properties of the edge-based frequent traversal graph for different k (horizontal axis).

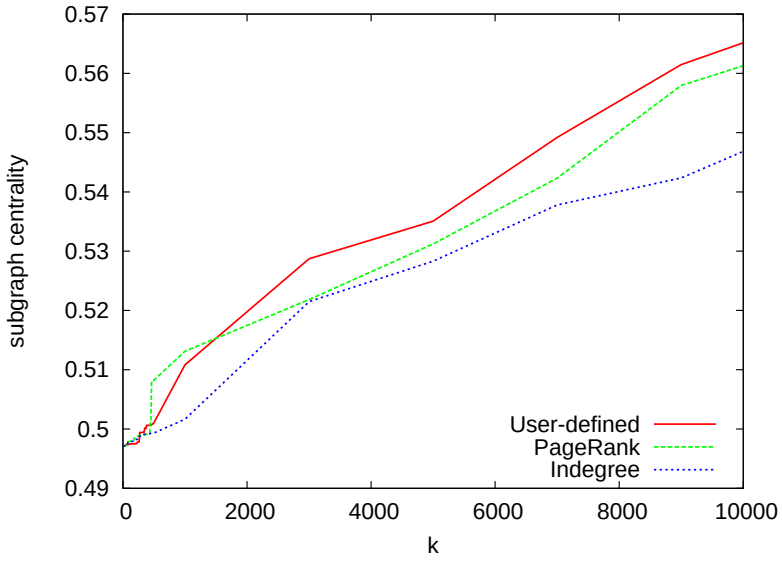

Figure 8.6: Comparison of subgraph centrality (vertical axis) of various centrality measures for different values of k (horizontal axis).

of the nodes of the graph which is better for realizing a low node-to-node distance than a traditional measure such as PageRank.

8.6 Conclusion

Throughout this chapter we have looked at mining path traversal patterns from the information network Wikipedia, aiming to understand and measure the quality in terms of navigation of user-generated traversal patterns. Using data gathered from over seven millions clicks made in the Wiki Game, we have derived a new measure of node centrality based on frequently traversed nodes.

It turns out that roughly half of the set of most frequently traversed nodes overlaps with the set of central nodes according to centrality measures such as PageRank. The additional nodes that are frequently visited by the users do appear to be useful, which we have demonstrated by using frequent traversal graphs and the notion of subgraph centrality. The subgraphs that can be derived from the frequently traversed nodes appear to be more central than the set of nodes derived from an existing centrality measure. This shows how users are apparently able to select an efficient portion of the graph that is useful in traversing the graph, specifically realizing a short distance to all other nodes in the graph. Although we have shown that the user is able to select

an efficient subset of the graph for completing navigation goals, it remains an open question exactly *how* the user selected this subset. Clearly, a subset derived using a centrality measure or a random subset performs similar or worse, so from an artificial intelligence point of view, the performance of the user is quite remarkable.

In future work, we want to see if we can extend the study of traversal patterns to more complex patterns based on frequent edges, possibly to study edge centrality [100], or based on frequent (interleaved) subpaths. Last but not least, the topics and techniques discussed in this chapter can possibly be extended to other types of graphs such as social networks, in which frequently traversed nodes and edges may indicate important actors and ties in the network.

Acknowledgment

We thank Alex Clemesha, creator of the Wiki Game, for providing the data.