Generalization in human pathfinding as a stochastic navigation process on Wikipedia

Dániel Ficzere (ficzeredaniel@vik.bme.hu)  
Budapest University of Technology and Economics

Gergely Hollósi (hollosi.ergely@vik.bme.hu)  
Budapest University of Technology and Economics

Attila Frankó (franko@tmit.bme.hu)  
Budapest University of Technology and Economics

András Gulyás (gulyas@gsuite.tmit.bme.hu)  
Budapest University of Technology and Economics

Research Article

Keywords:

DOI: https://doi.org/

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Generalization in human pathfinding as a stochastic navigation process on Wikipedia

Dániel Ficzere*, Gergely Hollósi, Attila E. Frankó and András Gulyás

Abstract

Models of human navigation have been investigated in many ways on complex networks. These findings suggest that the characteristics of human navigation change during navigation from the start to the destination. Nonetheless, the degree to which navigation is influenced by the human navigator versus the graph and environment remains somewhat unclear. Our study focuses on the initial stage of human navigation, investigating the influence of the graph structure on navigation through the use of a PageRank-based random walk model. We investigate the PageRank model with the classical uniform and also power-law position based transition probabilities. Our results indicate that a significant proportion of early generalization phase navigation can be simulated with random navigation.

Keywords: PageRank; Wikipedia; Human navigation; Random walk

Introduction

Navigation in complex networks is an important topic that has been investigated by researchers in a number of areas of computer science and beyond. Directed scale-free graphs contain nodes with high incoming degrees, also known as hubs, which appear to aid in navigating complex graphs. Research on human navigation in information networks suggests a two-phase process characterized by exploiting known information and exploring unknown information [1]. When individuals are confident about the links leading to their target, they follow those links purposefully (exploitation), but if they lack sufficient knowledge to relate candidate links to their goal, they select links almost randomly (exploration). During the exploration phase, individuals often visit high incoming degree nodes. Similar phase definitions are discussed in other studies, such as [2] and [3]. West et al. [3] reports that most participants navigate through high-degree hubs in the early phase, while content features guide their search thereafter. Lamprecht et al. [2] found a zoom-out and homing-in phase, with users guided by generality initially and textual similarity to the target later [4].

The main question is whether the early part – especially when the user is lost – can be described by random navigation, i.e. choosing the next edge by a specific probability distribution? In certain situations, such as being lost in a forest, humans tend to choose directions randomly until they stumble upon a path or trail. If this holds true for navigating through graphs, then the early stages are not solely dependent on human behavior but are instead an inherent property of the graph itself. Some researchers have used a node’s incoming degree as a measure of its
importance in finding a navigation path through a graph, as it may be correlated with the relative frequency of the node in navigational paths [2], [5]. To model random navigation and determine the relative frequency of nodes in this type of navigation, a first-order Markov chain is often used, assuming a randomly chosen next link from each node. However, calculating the stationary distribution for large graphs with millions of nodes can be computationally challenging, but the PageRank algorithm offers an efficient means of estimating it. The general PageRank model uses uniform transition probabilities between nodes. A more sophisticated approach to define the next link probability could be to use other factors such as link position, content similarity, topic similarity, and so on. Helic et al. [1] examines these factors in order to specify what makes a link successful on Wikipedia. The paper focuses on three types of link features, network features (indegree, outdegree, k-score), semantic similarity features, and visual features (x,y coordinates on the screen). Similarly to [1], we used the link position on a particular Wikipedia page to calculate the probability of the next page. However, we utilized the order of the link rather than the (x,y) coordinates on the screen as platform independent metric. Users tend to click next hops close to the top of the Wikipedia article; we formalize this behavior by modeling it with Zipf distribution.

To study random navigation in a complex graph, researchers have utilized the entire Wikipedia encyclopedia and the Wikigame goal-directed navigation game. Wikipedia is a vast information network with multiple hyperlinks to other topics, making it an ideal platform for evaluating human navigation behavior. Furthermore, Wikipedia’s node degree distribution follows a power-law, making it a scale-free network. In Wikigame, players are randomly assigned a start and destination article and must navigate between them using as few Wikipedia articles as possible. Players are not provided with any knowledge of the global network structure, relying solely on local information, such as outgoing links connecting the current article to its neighbors and their expectations about which articles are likely to be interlinked. As a result, Wikigame provides a reliable source of ground-truth human navigational patterns to compare with random navigation on Wikipedia.

As the main contribution of our paper, we define a models based on PageRank to describe the initial phase of human navigation on a scale-free network such as Wikipedia both for uniform and link position based distribution. The two PageRank models are based on either uniform or power-law relative position transition probabilities. To validate these models, we create the graph representation of Wikipedia and use the navigation of Wikigame users as a ground truth. We compared the results of the models to the indegree properties of the network. This research is an extended version of our previously published paper [6]. The main novelty compared to [6] is that we defined a new model – Zipf PageRank – using the power-law behavior based on the relative link position on a Wikipedia page. By observing that the real clicks follow a non-uniform distribution, we introduced the Zipf distribution and described how this model could be fit to real data with maximum likelihood estimation (MLE) using Bayesian optimization. We provide evidence on the Wikigame dataset to show why the model based on Zipf distribution is more accurate than the general uniform PageRank model. Finally, we validate this new model on three metrics compared to the uniform PageRank and the classic indegree model.
Methods

Pagerank

In order to describe random navigation, first-order discrete-time Markov chains with finite state space were used, where the transition probability distribution can be represented by the $P$ stochastic transition matrix, with the $(i, j)$th element defined by Eq. 1.

$$p_{ij} = Pr(X_{n+1} = j | X_n = i) \quad (1)$$

In a navigation graph, $p_{ij}$ is the probability of a node transition from node $i$ to node $j$, and $\sum_j p_{ij} = 1$. In an aperiodic, recurrent Markov-chain, the stationary distribution $\pi$ is a (row) vector, whose entries are non-negative and sum to 1 (i.e. $\sum_i \pi_i = 1$) and it satisfies Eq. 2 for a given $P$ transition matrix.

$$\pi P = \pi \quad (2)$$

As Eq. 2 shows, $\pi$ is a normalized multiple of a left eigenvector of the transition matrix $P$ with an eigenvalue of 1. Besides that, if the Markov chain is time-homogeneous also, $\pi$ can be calculated also by Eq. 3.

$$\lim_{k \to \infty} P^k = 1 \pi \quad (3)$$

In case of a navigation graph, the $\pi$ distribution gives the relative frequencies of the node visits after an infinite number of steps from a random starting article.

Supposing a large, asymmetric $P$ transition matrix, it is computationally challenging to calculate the stationary distribution of a Markov chain, e.g. in case of Wikipedia, $P$ is a matrix with 20 million rows and columns, for which the eigen decomposition is an unviable problem. The PageRank algorithm can help to solve such a problem and offers a good approximation for the stationary state of the Markov chain.

The PageRank algorithm was initially implemented in Google's search engine [7]. In PageRank, node’s importance can be interpreted as the more a node is pointed by important nodes, the more it is important. PageRank is equivalent to the stationary distribution of a random surfer following a memory-less Markov process. The PageRank algorithm defines the transition matrix from the $P$ matrix as

$$R = (1 - \epsilon)P + \frac{\epsilon}{N} 1_N, \quad (4)$$

where $\epsilon$ is the so-called damping factor, and $1_N$ is the matrix from ones with size $N \times N$. In the case of many real graph, the transition matrix is not ensured to be ergodic and the $\epsilon$ damping factor helps to make the Markov chain irreducible and
aperiodic. So during the random walk from node $j$, with probability $(1 - \epsilon)$ the user choose the next article uniformly with probability $p_{ij}$ or with probability $\epsilon$ the user teleport uniformly towards a arbitrary node of the network. These teleportations ensure that the user cannot be stuck in the network and that the steady state probability distribution is unique.

The stationary distribution is then calculated iteratively, starting from an $\pi^{(0)}$ initial distribution as

$$
\pi^{(t+1)} = \pi^{(t)} R,
$$

where the implementation of the PageRank algorithm can exploit the sparsity of matrix $P$. The convergence to the stationary distribution is governed by the second eigenvalue of the $R$ matrix, which is less than or equal to the $(1 - \epsilon)$ factor $[8]$. Thus, the error is decreasing with each step depending on the dumping factor as

$$
\text{Err}(n+1) \leq (1 - \epsilon) \text{Err}(n).
$$

By applying the recursive formula and a decent dumping factor, quick convergence can be reached to the stationary distribution. E.g., according to (6) the error decreases after 30 iterations with $\epsilon = 0.15$ by $(1 - 0.15)^{30}$, which is less then 1%.

**Uniform distribution**

Suppose some graph with $N$ nodes, with adjacency matrix $A$ and let $D = \text{diag}(d_1, \ldots, d_N)$, where $d_i = \sum_j a_{ij}$ is the outgoing degree of the node $i$. For all the nodes $i \in 1, \ldots, N$, if the outgoing transition probabilities has uniform distribution (i.e. $p_{ij} = 1/d_i$, also known as random walk), the transition probabilities can be calculated as

$$
P = D^{-1} A.
$$

For undirected graphs (i.e. $A = A^T$), the stationary distribution is proportional to the degrees of the nodes, meaning, that $\pi_i \propto d_i$ $[9]$. However, for directed graphs (i.e. $A \neq A^T$), no such simple, closed form solution can be found.

**Zipf distribution**

Examining the link click distribution of Wikigame clicks, it can be seen that the distribution of the transition probabilities for a node is not uniform, but clicks at the beginning of a page are more probable than clicks at the end of a page. Figure 1 presents the relative position histogram of the top 4 most used Wikipedia articles, while Figure 2 presents the relative position histogram from Wikigame. The latest 1 million clicks have been sampled from the Wikigame database. These figures clearly suggest that the uniform transition probability would not be an accurate model of the user navigation. Therefore, to model the distribution of clicks, one can utilize the Zipf’s law stating a power-law distribution. According to the law, the distribution of the click on a page can be written as

$$
p(k|\alpha, N) = \frac{1}{H_{N,\alpha}} k^{-\alpha} = \frac{k^{-\alpha}}{\sum_{n=1}^{N} n^{-\alpha}}
$$


where \( k = 1, \ldots, N \) is the link index, \( N \) is the number of links on the page, \( \alpha \) is the parameter of the distribution and \( H_{N,\alpha} = \sum_{n=1}^{N} n^{-\alpha} \) is the normalizing constant, called the \( N \)th generalized harmonic number. In figure 3 the Zipf Probability Mass Function is visualized for several different \( \alpha \) values.

![Relative position histograms of the top 4 Wikipedia pages from 1 million Wikigame player clicks.](image)

Figure 1: Relative position histograms of the top 4 Wikipedia pages from 1 million Wikigame player clicks. After filtering the latest 1 million clicks, we grouped by the clicks based on source articles and selected the top four articles.

**Estimating the parameter of the Zipf distribution**

To estimate the \( \alpha \) parameter in the distribution Eq. 8, maximum likelihood estimation can be applied. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.

Suppose a \( \Omega \) set of clicks, where \( \Omega = \{(k_i, N_i) \mid i = 1, \ldots, I\} \). The \( i \)th click is a pair of \((k_i, N_i)\), where \( k_i \) is the index of the clicked link and \( N_i \) is the number of links on the page for the \( i \)th click. We call the set \( \Omega \) as the observations for the
maximum likelihood estimation. Supposing independent clicks, we can express the likelihood function as

$$\mathcal{L}(\alpha; \Omega) = \prod_{i=1}^{I} \frac{k_i^{-\alpha}}{H_{N_i,\alpha}}$$

or the log-likelihood function

$$\ln \mathcal{L}(\alpha; \Omega) = \sum_{i=1}^{I} \ln \frac{k_i^{-\alpha}}{H_{N_i,\alpha}} = \sum_{i=1}^{I} \left( -\alpha \cdot \ln k_i - \ln \sum_{n=1}^{N_i} n^{-\alpha} \right)$$

(10)

From this, the maximum likelihood estimation of $\alpha$ is

$$\hat{\alpha} = \arg \max_{\alpha} \ln \mathcal{L}(\alpha; \Omega)$$

(11)

It is readily seen that the closed-form maximization of neither the likelihood nor the log-likelihood function is possible, mainly because of the summation in the second logarithm in the log-likelihood. Numerical optimization methods can be
applied, like Bayesian optimization to find the $\alpha$ which maximizes the likelihood function [10]. Bayesian optimization works by constructing a posterior distribution of functions (gaussian process) that best describes the function you want to optimize. As the number of observations grows, the posterior distribution improves, and the algorithm becomes more certain of which regions in parameter space are worth exploring.

**Evaluation**

We used the publicly available Wikipedia graph [11] and the well-known Wikigame dataset to examine random navigation on complex graphs. The entire English Wikipedia [1] was processed using its XML description, resulting in a graph representation of Wikipedia. We eliminated special cases, such as articles beginning with "Category:“, from the dataset. Additionally, we disregarded redirect links in the XML representation of Wikipedia, as these redirection steps are not visible to users during navigation. The resulting graph had more than 21 million nodes and 350 million edges.

We utilized a private dataset from Wikigame for our goal-directed navigation samples. The focus of our research is to examine and describe the average behavior of human navigation. To avoid any user biases or traits, we gathered 150,000 navigation paths from different users of the Wikigame dataset. We only selected samples from the basic game type, known as "speed-race," in which the objective is to reach the destination article from a random starting article with the least number of articles. We only included finished games in our collection, where the user’s path ends at the destination article. We excluded other game types like "5-click-to-Jesus" as it could significantly affect the results. It is crucial to consider game lengths as some navigation features can vary based on shorter or longer games. Figure 4 presents a normalized histogram of game lengths, indicating that over 99% of game lengths lie between 3 and 18 steps.

![Figure 4: Histogram of the game length in Wikigame and the calculated density function with 0.74 bandwidth value. It shows that a significant portion of game lengths are between 3 and 18 steps (more than 99%).](image)

[1]English dump, 2022. 01. 01.
We created a C++ application to process and calculate the PageRank values efficiently, while RocksDB was applied to store the dataset and for optimized lookups.

**Discussion**

At first glance, the zoom-out and homing-in phases can be identified on the indegree distribution of the nodes used in a step in Wikigame (Fig. 5). During the zoom-out phase (the generalization phase), indegrees increase, and then in the semantical phase (where the user navigates towards the destination node), indegrees decrease. However, the significance of indegree in graph navigation is not entirely clear, despite its importance in a directed complex graph. Conversely, the PR (PageRank) value (i.e. stationary probability) of nodes is more interpretable and meaningful in terms of probability. In this section, we analyze the behavior of human navigation using both indegree and uniform PR and Zipf PR values, focusing on the first generalization phase. Specifically, our hypothesis is that the graph structure plays a significant role in this phase of generalization, and that even random navigation could effectively model this phase by allowing a user to arrive at a "general" node.

![Figure 5: Wikipedia articles' indegree distribution per step for different Wikigame path lengths. The zoom-out and homing-in phases can be identified as in the zoom-out (generalization) phase the indegrees are increasing, then in the homing-in phase, the indegrees of the nodes are decreasing.](image)

**Maximizing the log-likelihood function**

We used Bayesian optimization to find the $\alpha$ which maximizes the likelihood function (Eq. 10). For the $\Omega$ observation set, the already mentioned latest 1 mil-
lion Wikigame clicks have been utilized The Bayesian optimization resulted an \( \alpha = 0.7885 \). The result of the log-likelihood function for different \( \alpha \) values is presented in Figure 6. One can see that the visual representation confirms the numerical MLE optimization result.

![Figure 6: Results of the log-likelihood function for different \( \alpha \) values calculated from 1 million Wikigame clicks.](image)

**Metrics**

To compare the evaluation results with the ground-truth values, we employed three different metrics. The first metric is known as the *page intersection ratio* (PIR), which measures the number of common pages among the top-\( N \) pages sorted separately by the uniform PR, Zipf PR the indegree, and the relative frequency of ground-truth usage values. To determine the relevance of this metric, we observed that a vast majority of clicks by Wikigame users come from a small portion of Wikipedia articles (as shown in Fig. 7), making the top articles based on Wikigame usage significant properties of navigation. Therefore, we used the top-\( N \) graph nodes in our metrics for further analysis of the user generalization problem.

![Figure 7: Complementary cumulative usage of Wikipedia articles by Wikigame users. It shows that 42.4% of all clicks comes from only 0.76% (1000 articles) of all used articles and 75.8% comes from 7.6% (10000 articles).](image)
The second metric we employed is the *histogram intersection ratio* (HIR) for the top-$N$ uniform PR, Zipf PR and indegree articles compared to the ground-truth usage values. As the top-$N$ used nodes of the Wikigame and the top-$N$ articles of indegree are not the same set, we analyzed the histograms based on the union of these two sets of article names. The uniform and Zipf PR articles underwent a similar evaluation process. After creating the normalized histograms, we analyzed the ”similarity” of the uniform PR, Zipf PR and indegree histograms compared to the ground-truth usage histogram with the HIR.

The final metric used is the *Jensen-Shannon divergence*, which measures the similarity between two probability distributions based on the *Kullback-Leibler divergence*. For discrete probability distributions $P$ and $Q$ defined on the same probability space, the Jensen-Shannon divergence is a symmetric metric (not like the Kullback-Leibler divergence) that always has a finite value, as defined by Eq. 12.

\[
D_{JS}(P \parallel Q) = \frac{1}{2}D_{KL}(P \parallel M) + \frac{1}{2}D_{KL}(Q \parallel M),
\]

where $M$ is defined by $M = \frac{1}{2}(P + Q)$, and $D_{KL}(P \parallel M)$ is the Kullback-Leibler divergence of the distributions $P$ and $M$.

**Model validation and comparison**

By utilizing the metrics mentioned earlier, we compared the indegree, the uniform PR and Zipf PR measures of the nodes with the actual navigation patterns in Wikigame. Although there is a certain degree of correlation between the models, there are still fluctuations and subtle distinctions as depicted in Figure 8. The most notable difference between Figure 8a and Figure 8b is that there are no small-indegree nodes in Figure 8b.

![Figure 8: Wikipedia indegrees of the top 1000 Wikipedia articles according to uniform PR and Zipf PR order.](image)
The PIR metric at different top-N pages is shown in Fig. 9. The figure indicates that sorting articles by the uniform PR and Zipf PR results in more overlap in the top-N used articles than the indegree at almost any N value. For instance, about 40% of the top 1000 used Wikigame articles are the same as the top 1000 highest indegree nodes, and approximately 50% of the top 1000 uniform and Zipf PR articles. Comparing the two PR metrics, the PIR is quite similar, but Zipf PR performs a bit better between the top 30 and top 1000 articles. Considering the vast number of articles in the Wikipedia evaluation (over 21 million), this fact is quite impressive. Additionally, the behavior is more precisely described by PageRank, indicating that random navigation leads to a behavior similar to user navigation.

![Figure 9: The intersection ratio of the top uniform PR, Zipf PR and top indegree articles compared to Wikigame usage for different number of top articles. It shows that the articles by the uniform PR and Zipf PR gives more overlap in the top-N used articles than the indegree at almost any N value.](image-url)

The nodes that are most frequently used are likely to be hubs in the network. Therefore, it is worth exploring different Wikigame step values to investigate the metrics. Our hypothesis is that, even unconsciously, users tend to navigate to hubs during the first phase of the game. Since we are focused on characterizing the generalization phase rather than the entire game, we analyze the PIR per step for three different numbers of top articles (50, 200, 1000), as shown in Figure 10. The results are generally similar to those in Figure 9, as the Zipf PR performs the best, the uniform PR follows as second and the indegree model have the lowest intersection ratios with the Wikigame. Additionally, it is important to note that the PIR ratio is highest during the first few steps and decreases quasi-monotonically.

Figure 11 presents the HIR metric, which indicates that the Zipf PR model has a higher histogram intersection ratio compared to the other two model. However, for the Top 1000 article scenario the uniform PR performs similarly well as the Zipf PR model. The ratio is significantly better for the first 2-3 steps of the game, which indicates different behavior from the users in the generalization phase of the game.
Figure 10: The intersection ratio of the top PageRank and top indegree articles compared to Wikigame usage for different number of top articles (50, 200, 1000) per step. It highlights that the PageRank model has higher intersection ratios with the Wikigame than the Indegree model. The metric is calculated for uniform distribution (---), for the Zipf distribution (- - -) and for the indegree (....).

Figure 11: The histogram intersection ratio of the top PageRank and top indegree articles compared to Wikigame usage for different number of top articles (50, 200, 1000) per step. First the union of the top 50/200/1000 articles were determined both for the PageRank–Wikigame usage and indegree–Wikigame usage, then the pdfs of the union set of articles have been calculated and compared as histogram intersection ratios. The metric is calculated for uniform distribution (---), for the Zipf distribution (- - -) and for the indegree (....). (Higher ratio means better fit.)

Furthermore, as shown in Figure 12, the results confirm that the probability distribution based on the Zipf PR model has a significantly better fit, as evidenced by the lower Jensen-Shannon divergence compared to the uniform PR and indegree model. This implies that the Zipf PR model is a more accurate description of the navigational process than the other two models. In conclusion, the main trends are quite similar to the previous cases.

Upon examination of the results, it is evident that the Wikigame has two distinct phases, namely the generalization phase and the home-in phase. The generalization phase is typically a brief part of the game that spans one or two steps, during which
users easily and quickly find hubs. Notably, the results suggest that hubs are reached with high probability when using random navigation on the Wikipedia graph, which indicates that the generalization phase is similar to random navigation, albeit not exactly the same.

Furthermore, the results reveal that the PageRank algorithm, which employs a well-interpretable random navigation method based on Markov chains, more accurately characterizes the generalization phase – especially the Zipf PR model – compared to the indegree distribution of the nodes. So, it can be stated that the relative position based transition probability model results in a more accurate model compared to the general uniform transition probability based models. Besides, PageRank also has the added benefit of considering the graph structure, as opposed to just the bare indegree of the nodes. However, random navigation is only well-applicable during the generalization phase. Once users reach a hub with a semantically connected link to the target node, they no longer behave randomly.

### Conclusion
The publicly available Wikipedia graph was utilized to investigate human navigation behavior in complex graphs in this study, where the ground-truth human navigation patterns were obtained from samples of the Wikigame application. The main inquiry focused on how human navigation patterns in the generalization phase differ from random navigation. Specifically, is human navigation during the generalization phase determined by human behavior or the graph structure?

Our study employed first-order Markov chains to confirm the conjecture that human navigation has two phases. Additionally, we demonstrated that the generalization phase is brief, with humans quickly identifying hubs. However, this rapidity...
is not exclusively due to human behavior, as the graph structure also has a significant impact on the generalization phase. We introduced a PageRank-based model – with Zipf and uniform transition probability distribution –, which outperformed the classical indegree based metric according to the evaluated metrics. Our study does not suggest that users solely engage in random navigation to generalize, but points out that a considerable proportion of their navigation behavior can be modeled using random navigation in complex graphs. It is well-know that random navigation and the structure of a scale-free graph results in high probability of reaching hubs in the graph – we state, that in case of missing navigational knowledge, humans deeply exploit this property of the graph of Wikipedia. In our view, random navigation by humans is not surprising; rather, it is a general behavior. When humans are lost in a navigation process, they attempt different directions to locate a familiar location from which they can continue their navigational purpose. Although semantical characteristics play a role in the navigation process, even in the first phase, our current research focuses on developing a comprehensive model that considers these characteristics. Additionally, we aim to examine and identify user-specific characteristics since this study only investigated average human navigational behavior.

Declarations

Funding
Not applicable.

Availability of data and materials
The Wikipedia dataset can be found https://meta.wikimedia.org/wiki/Data_dumps. For our work we used the 2022.01.01. Wikipedia dump.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors’ contributions
All authors contributed to create the main ideas behind the paper. D.F. and A. F. processed the dataset, while G.H. calculated the PageRank. D.F. calculated the metrics and created the figures. D.F. and G.H. wrote the main manuscript text. All authors reviewed and approved the manuscript.

Author details
Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary.

References


