Random Walk for Generalization in Goal-Directed Human Navigation on Wikipedia

Dániel Ficzere(✉), Gergely Hollósi, Attila Frankó, and András Gulyás

Budapest University of Technology and Economics, 1111 Műegyetem rkp. 3,
Budapest, Hungary
ficzere.daniel@vik.bme.hu

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 the navigation from the start to the destination. However, it is not fully clear to what extent the navigation is defined by the human navigator or the graph and the environment. Our work examines the early phase of human navigation, where we investigate the impact of the graph structure on human navigation with a random walk model based on PageRank. Our results suggest that a very high portion of human navigation in the early generalization phase can be modeled with random navigation.

Keywords: PageRank · Wikipedia · Human navigation · Random walk

1 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 have a couple of nodes with high incoming degrees (called hubs), which are seem to be responsible for navigating through complex graphs. It is observed e.g. that human navigation in information networks is a two-phase process combines of the phenomena they call the exploitation of the known and the exploration of the unknown [4]. Their results suggest that humans either follow specific links on purpose (exploitation) whenever they are confident enough that those links bring them closer to their particular target, or they select links almost arbitrarily at random (exploration), whenever they do not possess enough knowledge to relate the candidate links to their target. In the exploration part, the user often visits high incoming degree nodes in the graph. Similar phase definition are also discussed by [5] and [9]. [9] states that most subjects navigate through high-degree hubs in the early phase, while their search is guided by content features thereafter, while [5] mentions that their findings confirm that there is a zoom-out and a homing-in phase, where users are guided by generality at first and textual similarity to the target later [1].

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2016 2022, SCI 1077, pp. 202–213, 2023.
https://doi.org/10.1007/978-3-031-21127-0_17

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 uniform distribution? Human navigation can be random in certain circumstances, e.g. when lost in the forest, humans choose mainly random directions until finding some path or trail. If it is so, then the early part is not only about human behavior but is the *property* of the graph itself. A couple of works apply the incoming degree as the degree of importance of a node in finding the navigation path through the graph [2,5], i.e. it might be correlated with the relative frequency of the node in navigational paths. To model random navigation, and acquire the relative frequency of nodes in random navigation, a first-order Markov chain is applied, supposing a uniformly chosen next link from each node. However, while calculating the stationary distribution of a graph with a couple of million nodes is computationally hard; but the PageRank algorithm helps us to estimate it efficiently.

To investigate random navigation on a complex graph, the whole Wikipedia is used along with Wikigame[1] goal-directed navigation game. Wikipedia is the largest encyclopedia created ever. Besides the topic description, Wikipedia pages contain several hyperlinks to other topics, making Wikipedia an excellent information network for evaluating human navigation behavior. Furhtermore, Wikipedia is a scale-free network, so the node degree distribution follows a power law. In Wikigame, players randomly get a start and a destination article, and the goal is to solve the task of navigation from the start to the destination via as few Wikipedia articles as possible. Players have no knowledge of the global network structure besides semantical knowledge. Thus, they must rely solely on the local information—the outgoing links connecting the current article to its neighbors—and on their expectations about which articles are likely to be interlinked. Thus, Wikigame provides ground-truth human navigational patterns to be compared to random navigation on Wikipedia.

As the main contribution of our paper, we define a model based on PageRank to model the initial phase of human navigation on a scale-free network such as Wikipedia. To validate our model, we create the graph representation of Wikipedia and use the navigation of Wikigame users as a ground truth. We compared the results of our model to the indegree properties of the network as the indegree is also a good measure of generalization in a network. The paper is structured as follows: Sect. 2 introduces the Markov chains for random navigation, the role of node degree for stationary distributions and the applicability of PageRank algorithm for such a problem. The main properties of the used Wikipedia and Wikigame datasets are presented in Sect. 3. We introduce the key metrics and the model validation process for the generalization phase in human-goal directed navigation in Sect. 4. Section 5 concludes our paper and identify further research topics.

[1] https://www.thewikigame.com.

2 Methods

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** transition matrix, with the (i, j)th element defined by Eq. 1.

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

In a navigation graph, p_{ij} is the probability of a node transition from node i to node j. A stationary distribution π is a (row) vector, whose entries are non-negative and sum to 1 and it satisfies Eq. 2 for a given **P** transition matrix.

$$\pi \mathbf{P} = \pi \tag{2}$$

π is a normalized ($\sum_i \pi_i = 1$) multiple of a left eigenvector of the transition matrix P with an eigenvalue of 1. Besides that, the used Markov chain is time-homogeneous also, so the **P** can be calculated also by Eq. 3.

$$\lim_{k \to \infty} \mathbf{P}^k = \mathbf{1}\pi \tag{3}$$

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

Suppose some graph with N nodes, with adjacency matrix **A** and let $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_N)$, where $d_i = \sum_j a_{ij}$ is the outgoing degree of the node i. In case of a *random walk*, the transition probabilities can be calculated as

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}. \tag{4}$$

For undirected graphs (i.e. $\mathbf{A} = \mathbf{A}'$), the stationary distribution is proportional to the degrees of the nodes, meaning, that $\pi_i \propto d_i$ [6]. However, for directed graphs (i.e. $\mathbf{A} \neq \mathbf{A}'$), no such simple, closed form solution can be found.

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 [8]. 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

$$\mathbf{R} = (1 - \epsilon)\mathbf{P} + \frac{\epsilon}{N}\mathbf{1}_N, \tag{5}$$

where ϵ is the so-called damping factor, and $\mathbf{1}_N$ 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 ϵ 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 $1/d_j$ or with probability ϵ 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

$$\pi^{(t+1)} = \pi^{(t)}\mathbf{R}, \tag{6}$$

where the implementation of the PageRank algorithm can exploit the sparsity of matrix $\mathbf{P}$. The convergence to the stationary distribution is governed by the second eigenvalue of the $\mathbf{R}$ matrix, which is less than or equal to the $(1-\epsilon)$ factor [3]. Thus, the error is decreasing with each step depending on the dumping factor as

$$\mathrm{Err}(n+1) \leq (1-\epsilon)\mathrm{Err}(n). \tag{7}$$

By applying the recursive formula and a decent dumping factor, quick convergence can be reached to the stationary distribution. E.g., according to (7) the error decreases after 30 iterations with $\epsilon = 0.15$ by $(1-0.15)^{30}$, which is less then 1%.

3 Evaluation

To investigate the random navigation over complex graphs we used the publicly available Wikipedia [7] graph and the well-known Wikigame dataset for accessing the ground-truth human navigation patterns. The latest version of the whole English Wikipedia[2] have been processed and used, based on its XML description, the graph representation of the Wikipedia have been created. Special cases—such as articles starting with ,,Category:"—have been eliminated from the dataset. Also, there are redirect links in the XML Wikipedia representation, where the next hyperlink of an article redirects to another article. However, the redirection process is invisible for the user, therefore during navigation these redirection steps are also irrelevant. The final size of the graph consisted of more than 21 million nodes and 350 million edges.

For the goal-directed navigation samples we used the private dataset of Wikigame. The main purpose of our work is to analyze and characterize the average human navigation behavior, therefore we collected 150 000 navigation paths from the Wikigame dataset. Each navigation are produced by different users to eliminate user biases and characteristics. Also, there are multiple game types in Wikigame, so we sampled only from the basic game type, so-called "speed-race", where the main goal is to reach the destination article from a random start article via minimal number of articles. Besides that, only finished

[2] English dump, 2022.01.01.

games have been collected. A game is finished only when the destination article is the same as the last step of the user's path in Wikigame. These are quite important steps as the results could be very distorted if e.g. the "5-click-to-Jesus" game type is included to the analysis.

The normalized histogram of the game lengths (user navigation path lengths) is presented in Fig. 1. The figure depicts that more 99% of the game lengths are between 3 and 18 number of steps. In some cases it is important to consider the game lengths, because some characteristics of the navigation change observing shorter or longer games.

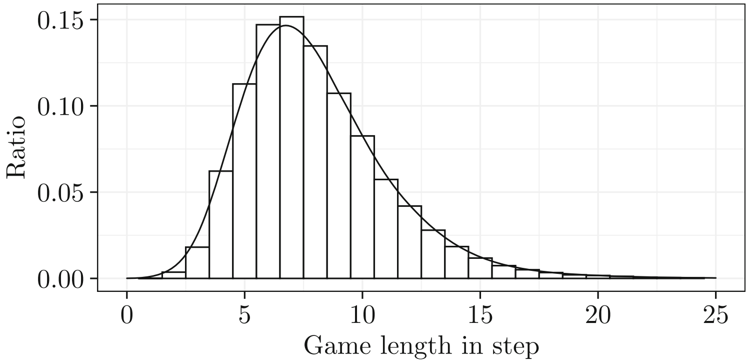
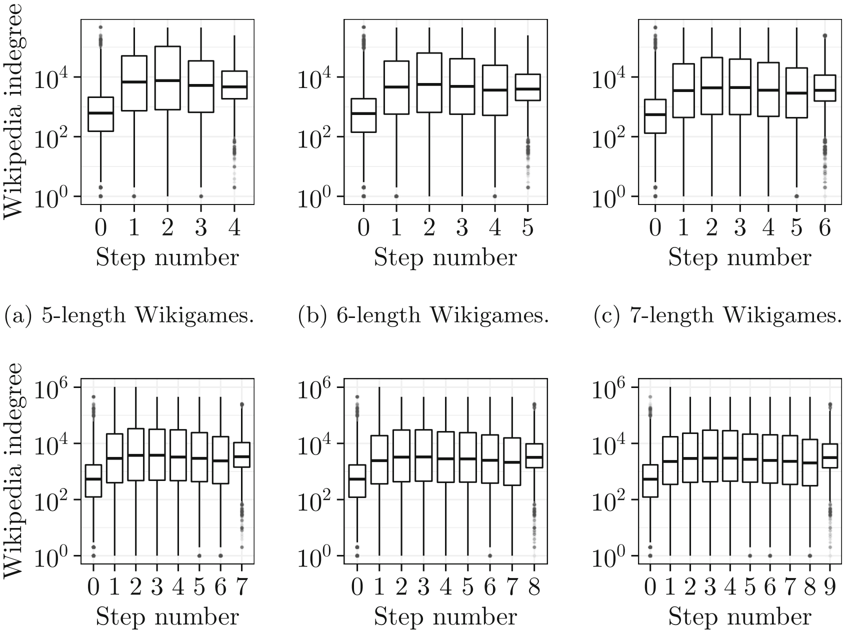


Fig. 1. Histogram of the game length in Wikigame and the calculated density function with 0.74 bandwidth value. It shows that a significant portion of game lengths are between 3 and 18 steps (more than 99%).

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.

4 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. 2). In the zoom-out phase (the generalization phase), the indegrees are increasing, then in the semantical phase (where the user tries to navigate towards the destination node), the indegrees of the nodes are decreasing. However, while the indegree of a node has a great importance in a directed complex graph, it is not clear, how to interpret the indegree regarding graph navigation. In contrast, the stationary distribution of a Markov chain has a well interpretable and probability-based meaning. In this section, we use both the indegree and the PageRank value (i.e. stationary probability) of the nodes to investigate the behavior of human navigation. Our analysis focused on the first generalization phase, more specifically, our hypothesis is that the graph structure has quite significant role in this generalization process, and even the random navigation would offer a great model to this phase, so a user could reach to a "general" node using even random navigation.



(a) 5-length Wikigames. (b) 6-length Wikigames. (c) 7-length Wikigames.

(d) 8-length Wikigames. (e) 9-length Wikigames. (f) 10-length Wikigames.

Fig. 2. 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.

4.1 Metrics

To compare ground-truth values and the result of the evaluation (the indegree and the PageRank), three different metrics were used.

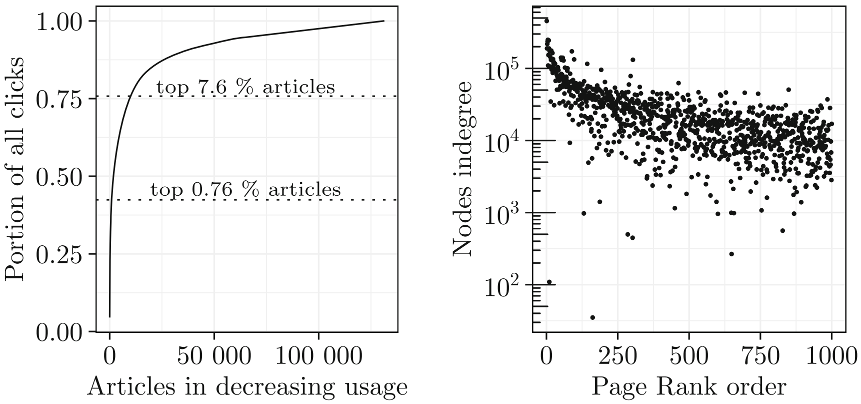
The first metric is the *page intersection ratio* (PIR), which is the number of common pages in the top-N pages sorted separately by the PageRank, the indegree and the relative-frequency of the ground-truth usage values. To establish the relevance of this metric, Fig. 3a shows that high portion of all click by Wikigame users comes from a very small portion of Wikipedia articles. Therefore the top articles based on Wikigame usage are quite characteristic properties of the navigation. Based on that we used the top-N graph nodes in our metrics during the further analysis of user generalization problem.

The second metric is the *histogram intersection ratio* (HIR) similarly for the top-N PageRank 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, the analyzed histograms are based on the union of this two set of article names. The evaluation process was similar for the PageRank articles also. So, after the creation of the normalized histograms we analyzed the ,,similarity" of the PageRank and indegree histograms compared to the ground-truth usage histogram with the HIR.

The last used metric is the *Jensen-Shannon divergence*, which is a method of measuring 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, χ, the Jensen-Shannon divergence is a symmetric metric and it always has a finite value as defined by Eq. 7.

$$D_{JS}(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M), \tag{8}$$

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.



(a) 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).

(b) Wikipedia indegrees of the top 1000 Wikipedia articles according to PankRank order.

Fig. 3. General Wikipedia statistics

4.2 Model Validation and Comparison

Using the metrics presented above, the two measures, i.e. the indegree and the PageRank of the nodes are compared to the ground-truth Wikigame navigation patterns. As it can be seen in Fig. 3b, the indegree and the pagerank are somewhat correlated, however, there are fluctuations and subtle differences, so the expected results are going to be different for the indegree and the PageRank.

Figure 4 shows the PIR metric at different N values in the top-N pages. As it can be seen, sorting the articles by the PageRank gives more overlap in the top-N used articles than the indegree at almost any N value. The figure shows that for example around 40% of the top 1000 used Wikigame articles are the same as the top 1000 highest indegree nodes and around 50% of the top 1000 Pagerank articles. In the context of how many articles are in the Wikipedia evaluation (more than 21 million), the fact is quite impressive. Also, PageRank seems to describe the behavior more precisely, meaning that random navigation results in a very similar behavior to user navigation.

However, the most used nodes are likely to be hubs in the network, so it's worth to investigate the metrics at different Wikigame step values, as our hypothesis is that users tend to navigate to hubs in the first phase of the game, even unconsciously. As we are trying to characterize the generalization phase not the whole game, we examine the PIR per step for three different number of top articles (50, 200, 1000) in Fig. 5. Overall the results are similar to Fig. 4, as the PageRank model has higher intersection ratios with the Wikigame, but another important fact is that the PIR ratio is the highest in the first few steps, and decreasing quasi-monotonously.

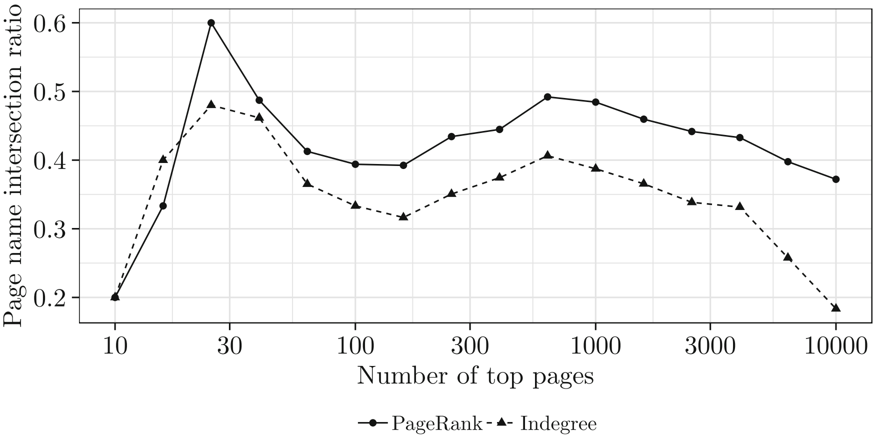


Fig. 4. The intersection ratio of the top PageRank and top indegree articles compared to Wikigame usage for different number of top articles. It shows that the articles by the PageRank gives more overlap in the top-N used articles than the indegree at almost any N value.

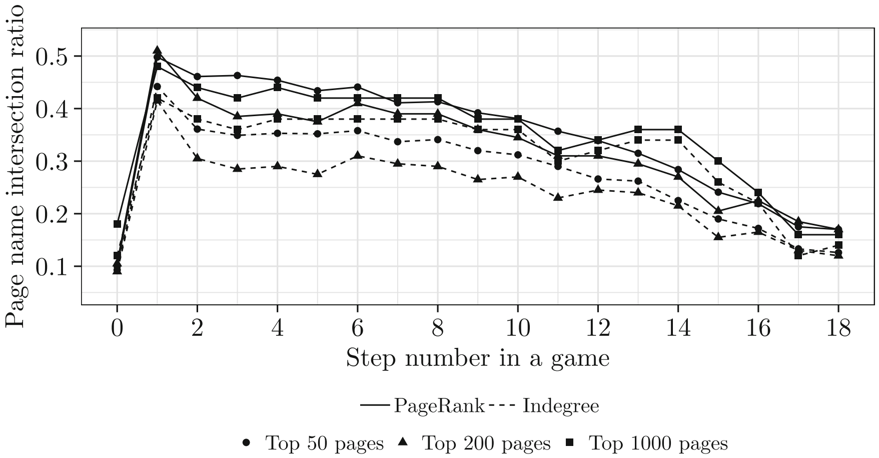


Fig. 5. The intersection ratio of the top PageRank and top indegree articles compared to Wikigame usage for different number of top articles (50, 200, 1000) per step. It highlights that the PageRank model has higher intersection ratios with the Wikigame than the Indegree model.

The HIR metric is presented in Fig. 6. The results show that the Pagerank model has a higher histogram intersection ratio for every examined cases, as well. 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 7 confirms the results above that the probability distribution based on the Pagerank model has better fit thanks to the lower Jensen-Shannon-divergence compared the indegree model. This means that the PageRank model describes better the navigational process than the indegree properties of the articles. The main trends are similar to the previous cases.

Looking at the results, it is clearly visible, that the game has basically two phases, the generalization phase and the home-in phase. The generalization phase is generally a short part of the game, consisting of 1 or 2 steps. Here, the user finds hubs very easily and quick. However, results show, that by applying random navigation on the Wikipedia graph, hubs are reached with high probability, so the generalization phase is really similar to a random navigation (yet not the same).

Also, the results show that the well interpretable random navigation based on Markov chains, and computed by the PageRank algorithm describes the generalization phase more precisely compared to the indegree distribution of the nodes. PageRank has also the advantage of taking the structure of the graph into account compared to the bare indegree number of the nodes. However, random navigation is applicable only to the generalization phase, when users reach a hub with a semantically connected link to the target node, it fails to behave randomly.

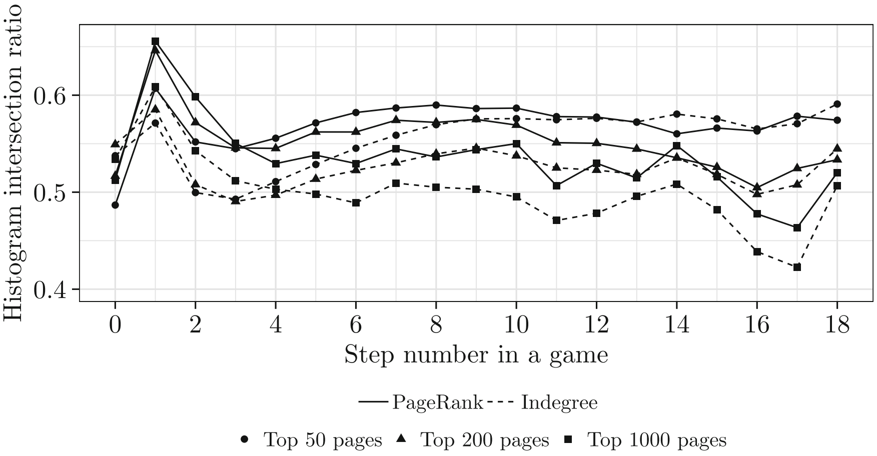


Fig. 6. The histogram intersection ratio of the top PageRank and top indegree articles compared to Wikigame usage for different number of top articles (50, 200, 1000) per step. First the union of the top 50/200/1000 articles were determined both for the PageRank–Wikigame usage and indegree–Wikigame usage, then the pdfs of the union set of articles have been calculated and compared as histogram intersection ratios. (Higher ratio means better fit.)

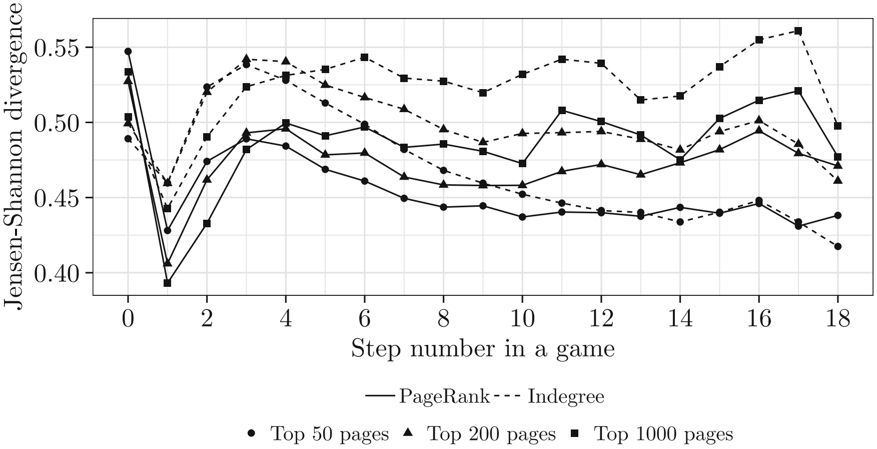


Fig. 7. The Jensen-Shannon divergence of the top PageRank and top indegree articles compared to Wikigame usage for different number of top articles (50, 200, 1000) per step. First the union of the top 50/200/1000 articles were determined both for the PageRank–Wikigame usage and indegree–Wikigame usage, then the pdfs of the union set of articles have been calculated and compared as Jensen-Shannon divergence. (Lower value means better fit.)

5 Conclusion

The paper investigated the human navigation behavior in complex graphs using the publicly available Wikipedia graph. The ground-truth human navigation patterns are samples from the Wikigame application. The main question was, that how different is the human navigation pattern in the generalization phase from random navigation? Or differently, is the human navigation determined in generalization phase by the human behavior, or by the graph structure?

In our work, the previously stated conjecture that the human navigation has at least two phases, was confirmed using first-order Markov-chains. Also, it was shown that the generalization phase is rather short and humans find hubs quickly. However, this quickness is not only a human behavior, the structure of the graph has a great impact on the generalization phase. We defined a model based on PageRank which has better properties according to the examined metrics than indegree which is also quite a good measure of generality. Our work do not state that users only navigate randomly to generalize, but a very high portion of their navigational behavior can be model with random navigation in complex graphs. In our opinion, random navigation by humans is not really surprising, moreover, it is a general behavior. When humans are lost in a navigation process (like in a forest or on an unknown street without a map), they try different directions to find some familiar place (handhold) from where they can continue the navigation purposeful. Obviously, semantical characteristics have a role in the navigation process even in the first phase, the creation of a comprehensive model where these characteristics are also considered is our current research topic. Moreover, we plan to investigate and identify user-specific characteristics as this paper only covers average human navigational behavior.

References

1. Berahmand, K., Nasiri, E., Forouzandeh, S., Li, Y.: A preference random walk algorithm for link prediction through mutual influence nodes in complex networks. J. King Saud Univ. Comput. Inf. Sci. (2021)
2. Gabrilovich, E., Markovitch, Shaul: Wikipedia-based semantic interpretation for natural language processing. J. Artif. Int. Res. **34**(1), 443–498 (2009)
3. Haveliwala, T., Kamvar, S.: The second eigenvalue of the google matrix. Technical Report 2003–20, Stanford InfoLab (2003)
4. Helic, D., Strohmaier, M., Granitzer, M., Scherer, R.: Models of human navigation in information networks based on decentralized search. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT '13, pp. 89–98, New York, NY, USA. Association for Computing Machinery (2013)
5. Lamprecht, Daniel, Lerman, Kristina, Helic, Denis, Strohmaier, Markus: How the structure of Wikipedia articles influences user navigation. New Rev. Hypermedia Multimed. **23**(1), 29–50 (2017)
6. Lovász, László.: Random walks on graphs. Combinatorics, Paul Erdos is eighty **2**(1–46), 4 (1993)
7. Meta-Wiki. Data dump torrents—Meta

8. Page, L., Brin, S., Motwani, R., Winograd, T.L: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November (1999). Previous number = SIDL-WP-1999-0120
9. West, R., Leskovec, J.: Human wayfinding in information networks. In: Proceedings of the 21st International Conference on World Wide Web, WWW '12, pp. 619–628, New York, NY, USA. Association for Computing Machinery (2012)