# Steering the Random Surfer on Directed Webgraphs

Florian Geigl
KTI, Graz University of Technology
Graz, Austria
Email: florian.geigl@tugraz.at

Simon Walk
IICM, Graz University of Technology
Graz, Austria
Email: swalk@iicm.edu

Markus Strohmaier
University of Koblenz-Landau and GESIS
Cologne, Germany
Email: markus.strohmaier@gesis.org

Denis Helic
KTI, Graz University of Technology
Graz, Austria
Email: dhelic@tugraz.at

***Abstract*—Ever since the inception of the Web website administrators have tried to steer user browsing behavior for a variety of reasons. For example, to be able to provide the most relevant information, for offering specific products, or to increase revenue from advertisements. One common approach to steer or bias the browsing behavior of users is to influence the link selection process by, for example, highlighting or repositioning links on a website. In this paper, we present a methodology for (i) expressing such *navigational biases* based on the *random surfer model*, and for (ii) measuring the consequences of the implemented biases. By adopting a model-based approach we are able to perform a wide range of experiments on seven empirical datasets. Our analyses allows us to gain novel insights into the consequences of navigational biases. Further, we unveil that navigational biases may have significant effects on the browsing processes of users and their typical whereabouts on a website. The first contribution of our work is the formalization of an approach to analyze consequences of navigational biases on the browsing dynamics and visit probabilities of specific pages of a website. Second, we apply this approach to analyze several empirical datasets and improve our understanding of the effects of different biases on real-world websites. In particular, we find that on webgraphs—contrary to undirected networks—typical biases always increase the certainty of the random surfer when selecting a link. Further, we observe significant side effects of biases, which indicate that for practical settings website administrators might need to carefully balance the desired outcomes against undesirable side effects.**

***Index Terms*—Navigational Biases, Biased Random Surfer, Stationary Distribution, Popularity**

## I. Introduction

Millions of people access the Web on a daily basis to conduct a variety of different tasks, such as maintaining social contacts, buying products in web shops, gathering information, or just passing time. While surfing the Web, users usually either traverse static (e.g., breadcrumb navigation) or dynamic (e.g., personalized recommendations) links, type in the URL of a website, or use a search engine to find their desired resource. Previous research has already established that users exhibit a $65\%$ probability of exploring websites through static links [10]. Many researchers already directed their efforts towards these $65\%$ of clicks, analyzing different aspects of the navigational behavior of users, such as estimating the probability of a user to traverse a given link by analyzing user click and interaction trails [13], [25]–[30]. Granka et al. [12] demonstrated how specific user behaviors directly influence which links are selected for browsing a website. Furthermore, Lerman and Ghosh [18] showed that users can be steered towards certain links by manipulating the interface (e.g., the position of links). In practice, website administrators often modify the interface to steer visitors towards certain pages. For example, owners of online shops might want to steer visitors towards best-sellers to increase revenue by modifying the probability that those pages are visited (e.g., by highlighting or repositioning links towards them).

**Problem & Approach.** Website administrators are typically not aware of the exact effects and implications of a particular modification. Moreover, such modifications may also affect the selection of other links and may trigger unpredictable and complex side effects. In fact, we still know very little about the (potentially) complex impacts of modifications and manipulations of linking structures on websites. In this paper we set out to close this knowledge gap. In particular, we aim at assisting website administrators in estimating the consequences of inducing specific biases on their website. In addition, we seek to increase our understanding of the emerging effects through biased link selection processes.

To this end, we present an approach for assessing the impact of different navigational biases on visit probabilities and browsing dynamics on directed webgraphs. We adopt a model-driven approach, based on the well-established random surfer model [3], to simulate users browsing a website. Although the model itself is very simple and straightforward, it provides a good approximation of actual user browsing behavior [7], [13], [29], [30].

In particular, we are interested in answering the following research questions:

*Website Coverage.* Can certain biases increase the effective number of pages visited by the random surfer or do they trap the surfer within specific (small) parts of a website?

*Surfer Guidance.* Given a specific bias, what is the degree of guidance (i.e., certainty) induced by that bias? How many options (on average) are random surfers confronted with when

they select the next link to follow? In other words, to what extent are browsing decisions purely random and to what extent do they adhere to a certain structure?

*Web Page Response.* How do visit probabilities of web pages respond to a given bias and how do such responses propagate through a network? For example, are those responses coupled and how? Specifically, what is the coupling between neighboring pages?

**Contributions.** In this paper we extend our framework[1] for simulating biased random surfers on networks [8] by analyzing, comparing and modeling the impact of unbiased and biased random surfers on directed webgraphs. In particular, we study real-world, empirical networks to obtain new insights into the global and local effects of different biases on the random surfer. Our results suggest that we can strongly and specifically influence the effective website coverage by using certain biases. Furthermore, we show that typical biases, such as popularity biases, always increase the certainty of the link selection process (i.e., provide a better guidance for the random surfer). Finally, we analyze potentially unwanted side effects that occur when inducing different biases, which affect a large proportion of all web pages of a website.

## II. Related Work

The random surfer is a simple but well-established model, which has already been extensively investigated by researchers in the past [19], [31]. It also represents the basis for the calculation of more complex node properties such as PageRank [3], [20] or HITS [16]. The PageRank model includes a parameter for the probability of the random surfer to teleport to a different node. This parameter is also often referred to as the damping factor $\alpha$, which is the probability that the random surfer continues to follow links at the current node. Conversely, with probability $1-\alpha$ the random surfer "jumps" to a randomly selected node and continues traversing links from there. Gleich et al. [10] empirically analyzed human click trails and estimated that the damping factor is in range between $0.6$ and $0.725$ for the Web.

Researchers have also manipulated the random surfer by applying different biases on the model to influence the link selection process [6], [11], [14], [22]. In such cases, the links are weighted and the link selection is not uniformly at random any more. Instead, the link selection probability is proportional to the link weights. Richardson et al. [23] used biased random surfers in the field of information retrieval. In their work they were able to outperform PageRank in terms of quality of the results. Despite an increase in computational costs and memory requirements, the authors argue that the algorithm is still reasonably feasible for large-scale search engines.

Al-Saffar and Heileman [1] later compared personalized and topic-sensitive PageRank with the original formula and came to the conclusion that both ways of personalization produce a considerable level of overlap in the top results. The authors conclude that new biases, which should not rely on the underlying link structure, are needed to improve the personalization of modified PageRank algorithms. In this paper we are not interested in improving the personalization of a node ranking algorithm. Instead we want to broaden our understanding of the effects of different biases on the stationary distribution of a random surfer.

West and Leskovec [30] investigated human click trails from a Wikipedia navigation game. Based on the results of this study, they [29] designed different features for steering a probabilistic random surfer. In their work they compared paths produced by the biased random surfer with those of humans. They found that human navigation was mostly based on popularity and similarity of articles. To further investigate this, we focus in this paper on the effects of popularity biases.

In 2013, Helic et al. [13] compared click trail characteristics of stochastically biased random surfers with those of humans. Their conclusion was that biased random surfers can serve as valid models of human navigation. In our previous work, we validated this finding by showing that the result vector of PageRank and click data biased PageRank have a strong correlation for the example of an online encyclopedia [7].

Regarding the number of pages which are effectively visited by random surfers, Hwang et al. [15] investigated the probability of returning to the start node of random surfers in scale-free networks. They found that this probability depends on the degree of the starting node, and thus the total distribution is similar to a power-law distribution. By investigating the stationary distribution of the random surfer, we circumvent this problem as the distribution is independent of the starting point.

In previous work [8] we have investigated how biases towards different subgroups of nodes influence the visit probability of the random surfer and how such biases compete with link insertion. In this paper we extend our methodology to allow for the simulation of biases based on structural properties of nodes, expanding the arsenal of tools to analyze the effects of biases on random surfers.

## III. Methodology

First, we introduce a basic notion for random surfers on a directed graph. Let $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of a directed graph with $W_{ij} = l$ where $l$ is the number of links that point from node $j$ to node $i$ (i.e., 0 if there are no links). The out-degree $k_i^+$ of a node $i$ is defined as the number of outgoing links, that is $k_i^+ = \sum_{j=1}^{n} W_{ji}$. Further, let $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix of weighted out-degrees (i.e., $d_{ii} = k_i^+$). Then the equation

$$\boldsymbol{P} = \boldsymbol{W}\boldsymbol{D}^{-1} \tag{1}$$

defines the transition matrix $\boldsymbol{P}$ with elements $P_{ij}$ equal to the probability of a random surfer moving from node $j$ to node $i$.

If we think about nodes as states and links as transitions between states, the transition matrix $\boldsymbol{P}$ defines a first-order Markov chain. If a Markov chain is irreducible (i.e., any

[1] The framework is available as open source software at https://github.com/floriangeigl/RandomSurfers

state can be reached from any other state with a non-zero probability) and aperiodic (i.e., returns to all states occur at irregular times), the chain has a unique stationary distribution $\boldsymbol{\pi}$. This distribution represents the probability of finding a random surfer on a given node in the limit of large number of steps. To ensure that the Markov chain $\boldsymbol{P}$ is irreducible we only use the largest strongly connected component from our datasets. On the other hand, a random walk on a connected directed graph is aperiodic if and only if there is no integer greater than 1 that divides the length of every cycle in the graph. Thus, it suffices to show that there is at least one cycle of length 2 and one cycle of length 3 in a directed graph for it to be aperiodic. We find that in all our datasets.

An algebraic solution for the stationary distribution yields $\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{\pi}$. Thus, the stationary distribution is an eigenvector of the transition matrix $\boldsymbol{P}$, corresponding to the largest eigenvalue 1. In related literature [2], [8], [17], the stationary probability of a node is often referred to as the *energy* of a node. As the random surfer is a conservative process [9], the system total energy is constant and equals 1. However, the distribution of energy over nodes is dependent on the link selection process of the random surfer under investigation.

**Inducing Bias.** In practice, we can influence the link selection process of users by, for example, repositioning links [18]. In our analysis, we bias the random surfer by weighting links in a given network to achieve similar effects. To that end, we investigate different structural properties of nodes and weight all links pointing towards nodes proportional to a given structural property. For example, to induce a *popularity* bias we weight links according to the popularity (i.e., degree) of the target node.

Algebraically, we represent a bias as a diagonal matrix $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ with node weights $\boldsymbol{b} \in \mathbb{R}^n$ on its diagonal. Matrix $\boldsymbol{W'}$ is then the weighted adjacency matrix of the biased network, which we calculate as the product of $\boldsymbol{B}$ and $\boldsymbol{W}$:

$$\boldsymbol{W'} = \boldsymbol{B}\boldsymbol{W}\,. \tag{2}$$

Using the weighted out-degree diagonal matrix $\boldsymbol{D'}$ of $\boldsymbol{W'}$ we calculate the corresponding biased transition matrix $\boldsymbol{P'}$ as:

$$\boldsymbol{P'} = \boldsymbol{W'}\boldsymbol{D'}^{-1}\,. \tag{3}$$

As before, we have the stationary distribution satisfying the right eigenvector equation (i.e., $\boldsymbol{\pi'} = \boldsymbol{P'}\boldsymbol{\pi'}$), where we use $\boldsymbol{\pi'}$ to denote the stationary distribution of the biased random surfer. Note that this methodology adapts and extends our previous work [8]. However, in this paper we do not bias towards groups of nodes but rather set the probability of traversing a link proportional to structural properties of its target node. Hence, all links of the network are affected by the induced bias as opposed to our previous work, where only links pointing towards selected nodes were affected. In practice this would mean that we highlight each link proportional to a property of the target page (e.g., popularity).

## IV. Experimental Setup

**Website Coverage.** In general, biases allow us to manipulate the link selection process of random surfers and influence the visit probabilities of specific nodes. To investigate the bias effects on the *effective* number of visited pages (i.e., pages with practically relevant visit probability) we calculate three properties of the stationary distribution. First, we analyze the visit probability of the most visited page of each website to see and compare how likely the random surfer can be found on just this single page. In all our datasets we find that the most visited page is always the *home page* (i.e., main/entry page) of the website. Second, we use the complementary cumulative distribution function of the stationary distribution (i.e., $CCDF(\boldsymbol{\pi})$) to determine the number of pages on which the random surfer can be found with a probability higher than 95%. Third, we analyze the entropy of the stationary distribution $H$, which measures the uncertainty in the current location of the random surfer. We calculate $H$ as:

$$H = -\sum_i \pi_i \log_2 \pi_i\,. \tag{4}$$

**Surfer Guidance.** To analyze the dynamics of the link selection process we calculate the entropy rate of the random surfer. Entropy rate is the average entropy of all decisions made by the random surfer in the limit of a large number of steps. Thus, it measures average uncertainty in all the decisions made by a random surfer. We calculate the entropy rate $H_{rate}$ as:

$$H_{rate} = -\sum_{ij} \pi_j P_{ij} \log_2 P_{ij}\,. \tag{5}$$

Note that the entropy of each node is weighted with the corresponding value of the stationary distribution. Thus, the uncertainty of the random surfer at a highly visited page has a greater impact on the entropy rate than the one from a less frequently visited page.

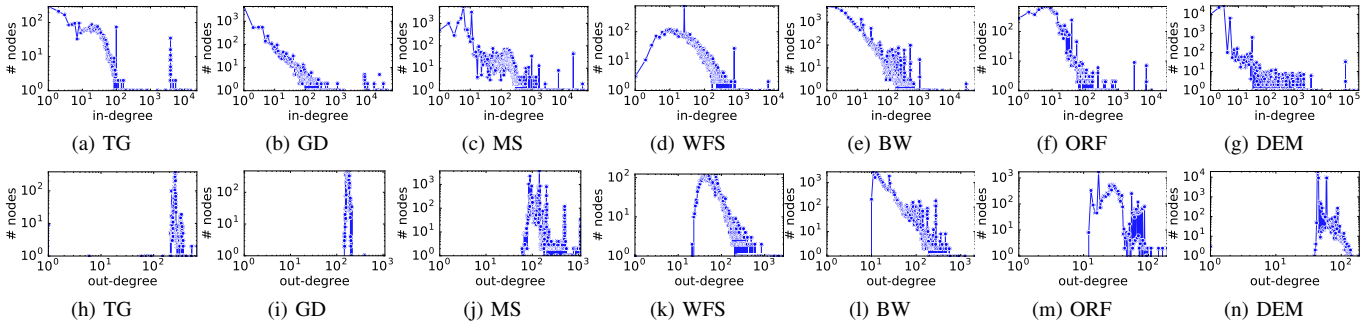
**Web Page Response.** To improve our understanding of changes in the visit probabilities of the random surfer due to different biases, we investigate how each individual page is affected on a microscopic level. We do that by analyzing heat maps which are based on log-scaled scatter plots between stationary distributions of unbiased and biased random surfers.

## V. Biases

In this section we introduce the investigated biases and the intuitions behind them.

**Popularity Bias.** With this bias we steer the random surfer towards popular nodes. For example, in web shops operators may want to increase visits (and thus potentially sales) of frequently visited products. In encyclopedias and media libraries, operators may have an interest in further increasing the visibility of popular articles or movies. For popularity we use the *degree* of web pages as a proxy and set $b_i = k_i$, where $k_i$ is the total degree (in and out) of node $i$.

**Unpopularity Bias.** To dampen the natural attraction of popular nodes we may want to induce an unpopularity bias. As a web shop operator, this could be used in a strategy to clear out stocks by increasing the visibility of unpopular items. In encyclopedias or media libraries operators may want to ease-up and diversify navigation to specific (mostly




(a) TG (b) GD (c) MS (d) WFS (e) BW (f) ORF (g) DEM

(h) TG (i) GD (j) MS (k) WFS (l) BW (m) ORF (n) DEM

Fig. 1: **In-Degree and Out-Degree Distributions of our Datasets.** The Figure depicts the in-degree (*top*) and out-degree (*bottom*) distributions of all datasets. The in-degree distributions are skewed towards few pages with a very large in-degree, which is typical for webgraphs. In contrast, the out-degree distributions are more homogeneously distributed, except for the Wikipedia datasets (*WFS*, *BW*). This is due to the way the websites are designed. In webshops (*TG*, *GD* and *MS*) and online media libraries (*ORF* and *DEM*) most pages are similarly structured and thus contain roughly the same number of outgoing links. On Wikipedia pages widely vary in their length, which is why the out-degree varies more strongly.

unpopular) pages and decrease the visibility of popular pages. For unpopularity bias we use the *inverse degree* and set $b_i = {}^1/_{k_i}$ for all $i$.

**Eigenvector centrality.** A bias proportional to eigenvector centrality of a node has already been investigated by researchers on *unweighted*, *undirected* networks [4], [5], [21], [24]. In such networks the eigenvector centrality bias produces the highest possible entropy rate [24]. Therefore, we include this bias in our experiments as a baseline. Eigenvector centrality is the right eigenvector of the adjacency matrix of a network and satisfies $\boldsymbol{W}\boldsymbol{v} = \kappa_1 \boldsymbol{v}$, where $\boldsymbol{W}$ is the weighted adjacency matrix of the network and $\kappa_1$ the largest eigenvalue of $\boldsymbol{W}$. Thus, we introduce the eigenvector centrality bias by setting $b_i = v_i$ for all $i$.

## VI. Datasets

To simulate navigational biases "in the wild", we have crawled webgraphs of seven different websites. In particular, we collected data from three web shops that deal with "geeky" gadgets or board games (ThinkGeek[2], GetDigital[3] and Milan-Spiele[4]), two online encyclopedias (Wikipedia for Schools[5] and Bavarian Wikipedia[6]), as well as two online media libraries (ORF TVThek[7] and Das Erste Mediathek[8]). In the remainder of the paper we will refer to the datasets using the abbreviations of their names denoted in Table I. The degree distribution of all datasets are depicted in Figure 1.

Concerning the crawling process itself, our web crawler recursively extracted and followed all links, starting from the main page of each website. Note that we did not fully render each page individually, resulting in the omission of links generated via (client-rendered) AJAX queries and Flash content. In a post-processing step we have removed self-loops—links from a web page to itself. Further, we preprocessed and removed links that coincide with several different redundant actions, such as links containing *?sessid=* or *?oCsid=* for session identifiers, *action=review* for displaying the "write a review" box, as well as "add to shopping cart", or "log-in" personalized user account links and parameters. From the cleaned datasets we constructed the corresponding webgraphs.

For the actual simulations we extracted the largest strongly connected component of the network (i.e., the largest subset of nodes in which every node can be reached from all other nodes) so that the random surfer does not get stuck on pages without outgoing links.

[2] http://www.thinkgeek.com
[3] http://www.getdigital.eu
[4] http://www.milan-spiele.de
[5] http://schools-wikipedia.org/
[6] https://bar.wikipedia.org
[7] http://tvthek.orf.at/
[8] http://mediathek.daserste.de/

TABLE I: **Network Statistics.** The table displays the basic statistics of our datasets, with $n$ being the number of nodes, $m$ the number of edges, and $d$ the network diameter.

| Dataset | Category | $n$ | $m$ | $d$ |
|---|---|---|---|---|
| **ThinkGeek (TG)** | web shop | 3,884 | 1,002,226 | 3 |
| **GetDigital (GD)** | web shop | 8,258 | 2,101,254 | 21 |
| **Milan-Spiele (MS)** | web shop | 21,566 | 3,128,693 | 70 |
| **Wikipedia for Schools (WFS)** | encyclopedia | 6,796 | 646,646 | 4 |
| **Bavarian Wikipedia (BW)** | encyclopedia | 32,734 | 1,324,839 | 9 |
| **ORF TVThek (ORF)** | media library | 9,799 | 301,844 | 10 |
| **Das Erste Mediathek (DEM)** | media library | 70,063 | 3,448,513 | 2274 |

## VII. Results & Discussion

### A. Website Coverage

The left part of the Table II depicts the results for *Website Coverage*. For almost all datasets the popularity biased random surfer achieves (i) the highest probability of being on the home page, (ii) the lowest number of nodes needed to reach an aggregated energy of 95% and (iii) the lowest stationary entropy. These results indicate a low website coverage, meaning that with a high probability we will find the random surfer on just a few nodes of the network. In other words, the random surfer is trapped on just a few pages of the website. On the other hand, we observe the opposite behavior for the unpopularity biased random surfer (cf. Table II). These results follow our intuition. We expect that in a network with an initially skewed stationary distribution, where just a few top nodes possess

TABLE II: **Website Coverage and Surfer Guidance.** Table II depicts the results of our experiments for all biases (columns) and all datasets (rows). The highest values for each dataset in each of the four sections (i.e., Home page, 95%, Stationary Entropy, and Entropy Rate) are highlighted in blue, and the lowest are marked in red. All three *Website Coverage* measurements indicate that a popularity bias (pop.) decreases website coverage, whereas the unpopularity bias (unpop.) is able to increase it. Hence, a bias towards popular pages traps the random surfer within a few pages, while the unpopularity bias allows random surfers to effectively visit more pages. The *Surfer Guidance* is represented by the *Entropy Rate*, which is the uncertainty of the random surfer when selecting a link to traverse. All biases are able to increase the certainty of the random surfer.Furthermore, eigenvector centrality biases (e.c.) in directed networks do not produce the highest entropy rate in our datasets, which is caused by the weak correlation between nodes in- and out-degrees.

| | **Website Coverage** | | | | | | | | | **Surfer Guidance** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Home page* | | | *95%* | | | *Stationary Entropy H* | | | *Entropy Rate* $H_{rate}$ | | | |
| | unb. | pop. | unpop. | unb. | pop. | unpop. | unb. | pop. | unpop. | unb. | pop. | unpop. | e.c. |
| TG | 0.01% | 0.00% | 0.01% | 1547 (39.83%) | 184 (4.74%) | 2431 (62.59%) | 8.78 | 7.16 | 10.66 | 7.64 | 6.86 | 6.41 | 6.76 |
| GD | 2.88% | 5.82% | 0.29% | 177 (2.09%) | 68 (0.80%) | 1165 (13.74%) | 6.86 | 5.40 | 9.93 | 6.38 | 5.31 | 5.01 | 5.35 |
| MS | 1.00% | 1.17% | 0.13% | 599 (2.78%) | 65 (0.30%) | 4082 (18.93%) | 7.94 | 5.99 | 11.32 | 6.11 | 5.61 | 4.02 | 5.67 |
| WFS | 3.13% | 13.04% | 0.09% | 3229 (47.51%) | 22 (0.32%) | 4348 (63.98%) | 9.65 | 4.79 | 12.01 | 5.61 | 4.24 | 4.68 | 4.16 |
| BW | 5.59% | 21.62% | 0.05% | 4563 (13.94%) | 23 (0.07%) | 14751 (45.06%) | 9.28 | 4.05 | 13.54 | 4.98 | 3.49 | 2.61 | 3.51 |
| ORF | 12.06% | 36.00% | 0.13% | 1398 (14.27%) | 11 (0.11%) | 3321 (33.89%) | 7.56 | 3.25 | 11.04 | 4.76 | 2.83 | 3.03 | 3.61 |
| DEM | 1.41% | 1.94% | 0.03% | 446 (0.64%) | 38 (0.05%) | 2812 (4.01%) | 7.55 | 5.11 | 10.75 | 5.60 | 4.63 | 1.94 | 5.15 |

an aggregated energy of almost 1, adding an additional bias towards these top nodes increases the skewness. The more skewed the distribution becomes, the likelier the random surfer gets trapped within the most popular nodes. On the other hand, a bias towards less popular nodes reduces the skewness of the stationary distribution.

**Findings & Implications.** The popularity bias decreases the website coverage, whereas the unpopularity bias is able to increase it. To raise the effective website coverage we should counteract the natural skewness of the stationary distribution of a webgraph. We can achieve this by, for example, inducing an unpopularity bias. Such a bias makes particularly sense in the case of online encyclopedias, where users should be able to easily explore the whole content of the website. However, in cases in which website administrators want to reduce costs (e.g., keep just a few movies on expensive, fast accessible storage devices in media libraries such as *ORF* or *DEM*), a bias towards popular web pages represents a suitable method.

### *B. Surfer Guidance*

The second column of Table II (Surfer Guidance) depicts the entropy rate of all combinations of datasets and biases. We find that the unbiased random surfer consistently exhibits the highest entropy rate across all datasets. This means that the guidance (i.e., certainty in link-selection decisions) of this surfer is low. Note that this is not the case in undirected, unweighted networks, where the eigenvector centrality bias generates the highest (maximum) entropy rates. Random surfers biased by eigenvector centrality exhibit similar entropy rates to the ones biased by popularity across all datasets except for *ORF* and *DEM*. Across all tested biases we achieve the lowest entropy rate for almost all datasets with the unpopularity bias. As steering the random surfer towards unpopular nodes decreases the average number of possible next hops a lower entropy rate is to be expected. However, for *WFS* the eigenvector centrality bias and for *ORF* the degree bias result in the lowest entropy rates.

Both effects—lowest entropy rate for one of the two biases towards popular nodes in *WFS* and *ORF* and the unobserved maximum entropy rate of the eigenvector centrality bias—are caused by specifics of webgraphs topologies. In particular, in our datasets we do not observe a strong correlation between in-degree and out-degree of a node. A possible explanation for this behavior is based on a specific information architecture on the Web and usability considerations. More specifically, websites tend to have a few pages with many incoming links. For example, on many websites all pages contain the website logo on the top, which is linked to the home page. In all our datasets we confirm this assumption by measuring an unweighted in-degree of $n-1$ for the home page, where $n$ is the number of pages of a website. On the other hand, the majority of other pages have only a few incoming links. Thus, there is a high variability in the number of incoming links. On contrary, the number of outgoing links is much more stable and in a typical cases limited due to usability reasons.

As a consequence of such network topology, the unbiased random surfer often visits nodes with a high in-degree. However, these nodes are often not the ones with the highest out-degree (e.g., the home page of websites often contains only very few outgoing links towards other very popular pages). Consequently, the random surfer has to choose between a few links only. This in turn keeps the uncertainty and the entropy rate low.

Note that in the case of undirected networks the random surfer often visits high-degree nodes, which bear decisions with the highest number of possible outcomes, resulting in highest entropy rates. To find further evidence in favor of our hypothesis we biased the random surfer with node out-degree. This experiment resulted in entropy rates higher than the one of the unbiased random surfer in most datasets.

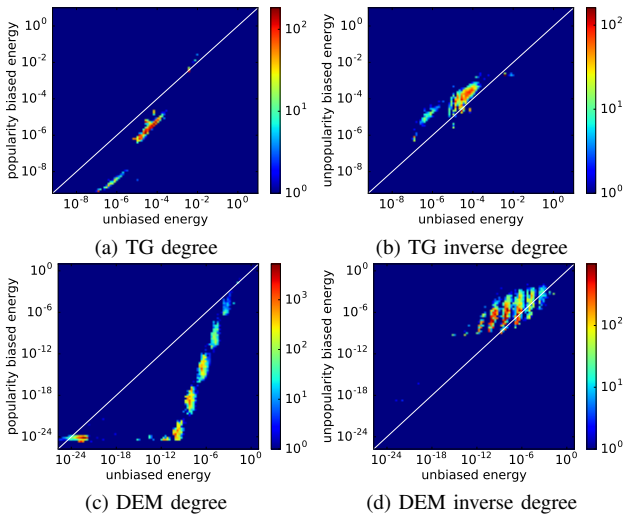
**Findings & Implications.** Both popularity and unpopularity bias reduce the entropy rate and at the same time increase certainty for random surfers on directed webgraphs. Consequently, both biases can serve as a way to increase the guidance throughout the website. This finding is in a stark contrast to undirected networks where the popularity biases increase the entropy rate.

### C. Web Page Response

In this experiment we investigate the response of individual pages to a bias. In the case of the popularity bias the majority of pages yields a part of their energy to just a few top pages (see Figure 2a and 2c). On the other hand, in the case of the unpopularity bias we observe a flow of energy from the top pages towards pages with a low initial energy (see Figure 2b and 2d). For the popularity bias we observe a slightly left-oriented v-shape in the scatter plots for some datasets (i.e., *MS* and *DEM*). This observation is particularly pronounced for *DEM* (Figure 2c). In general, this means that pages with a low initial energy (which are typically more distant to the top pages) are less affected (relatively) by the bias than, for example, pages with an average initial energy (which are closer to the top pages). Note that we can only observe such v-shapes for datasets with a very high (pseudo) diameter (cf. Table I).

In Figure 3a we plot the initial stationary distribution against the one from the popularity biased random surfer. We mark top pages as pages with an increased energy due to the induced popularity bias (see *top* in Figure 3a). We then color nodes based on their shortest distance to any of these top pages. The figure shows that these distances have a decisive effect on the biased page energy. A possible explanation for this behavior is that the top pages exhaust energy from all other pages, that is, the energy flows from all other pages towards top pages. The strength of the flow seems to be inversely proportional to the distance from the top nodes—the further away pages are from the top pages, the smaller the flow of energy towards top nodes. After a certain distance (i.e., 6 for *DEM* in Figure 3a) some pages reach the lowest possible state of energy and fall into a ground state (see *ground state* in Figure 3a). A page in this ground state practically loses all of its energy and thus its visit probability. The pages depicted around the low circle in Figure 3b were able to minimally increase their energy ($\approx 10^{-20}$ in *DEM*). We attribute this negligible effect to numerical inaccuracies.

To further analyze the energy flow in a network we group all nodes according to their shortest distance from the top nodes and calculate the energy as a function of this distance. We introduce two new popularity biases. First a bias proportional to square degree and second a bias proportional to square root degree of a node. We assume that the flow of energy towards top nodes must be the fastest in the case of the square degree, followed by the degree bias and then by the square root degree. In Figure 3b we see that all popularity biases concentrate the energy on just a few nodes and hence result in most of the other nodes falling into the ground state. We are also able to confirm our energy flow assumption since the flow



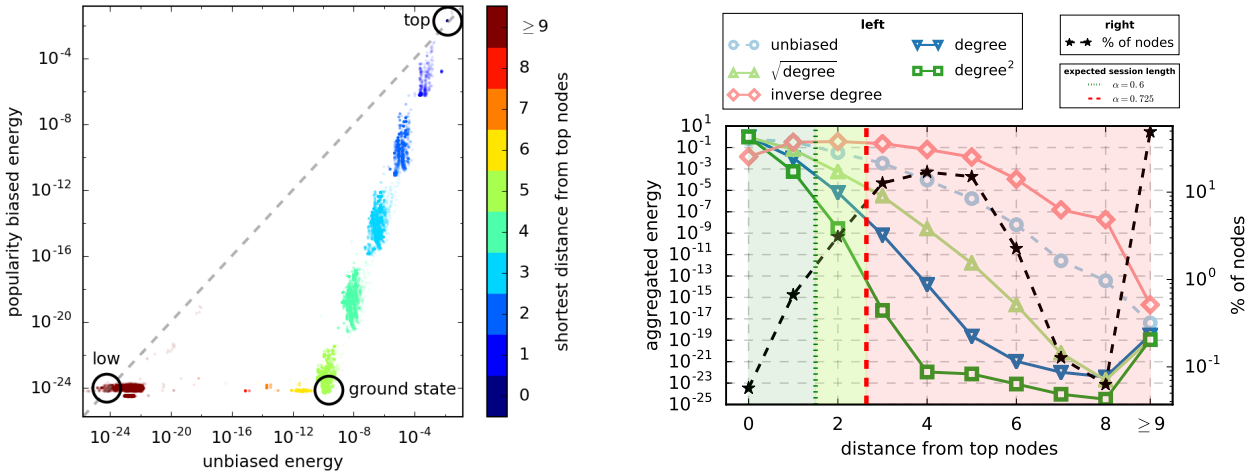

(a) TG degree (b) TG inverse degree

(c) DEM degree (d) DEM inverse degree

Fig. 2: **Web Page Response.** The heat maps depict the absolute energy gains and losses of all nodes due to an induced bias. The $x$-axes correspond to the unbiased energy of a node, whereas the y-axes denote the biased energy. Color refers to the number of nodes observed in that area. The white dashed diagonal marks perfect correlation (i.e., the energy of nodes on this line did not change). For the *TG* dataset (*top*) all biases result in the expected change of the stationary distribution. A popularity bias increases the energy of popular nodes while it decreases the energy of all other nodes. The opposite is true for the unpopularity bias. However, in some datasets, such as *DEM* (*bottom row*), we can see a slightly left-oriented v-shape, where nodes with average initial energies are most affected (relatively) by the bias in the form of decreased energy.

is strongest for the square degree (the nodes at distance 4 fall into the ground state), followed by the degree bias (the ground state is reached at distance 6). The square root and inverse degree distribute the energy more uniformly over distances (the ground state is reached at distance 8).

To get a better understanding of practical implications of these results we also calculate the expected browsing session length using empirically measured damping factors (i.e., $0.6 \leq \alpha \leq 0.725$ [10]). In particular, we can model session length as a random variable following geometric distribution with the parameter $1 - \alpha$. The expected session length equals then to $\alpha/(1-\alpha)$. Using empirical damping factors the expected session length lies between 1.5 and 2.64 clicks. Assuming that users start browsing on the home page of a website, the pages that they are expected to visit are within the distance of the expected session length. We mark the range of the expected session length as vertical lines in Figure 3b. Only pages that are at distance shorter then the expected session length and have a practically relevant stationary probability can be visited by users. Pages that are further away or are close but are fallen to the ground state will not be visited. Thus, in practice we may be able to increase the visibility of, for example, popular pages but because of a fast energy flow many other pages




(a) Degree Bias

(b) Distance Energy Level

Fig. 3: **DEM Energy Concentration.** These plots describe how energy diffuses when a bias is induced. The *left* plot depicts the stationary distribution of the unbiased random surfer ($x$-axis) against the one of the popularity biased random surfer ($y$-axis). We mark pages which increased their energy due to the bias as *top* pages and color each page based on its shortest distance from any of the *top* pages. All pages further away than 6 hops from top pages slide into the ground state of the system—meaning that they will almost never be visited by the random surfer. The minimal increase in energy ($\approx 10^{-20}$) of pages marked as *low* is likely caused by numerical inaccuracies. The *right* plot depicts groups of pages based on their shortest distance from top pages (x-axis) and their aggregated energy (left y-axis). Additionally, we show the fraction of pages (right y-axis; dashed black line with stars) for each distance and the range of expected sessions lengths based on the damping factors $\alpha$. The colored areas refer to the probability of users reaching nodes of a certain distance, if they start navigating from the home page (i.e., green: very likely, yellow: likely and red: unlikely). We see that, due to the unpopularity bias, a large amount of energy diffuses towards pages being 2 to 3 hops away from top pages. In contrast, all popularity biases (i.e., degree, $\sqrt{\text{degree}}$ and degree$^2$) concentrate the energy on just a few pages while pushing many other pages into the ground state. This means that small increases in energy of the *top* pages lead to many other pages being pushed into the ground state.

will practically become invisible to users. Therefore, biasing link selection process includes a trade off between desirable outcomes and (possibly) unwanted side effects.

Specifically, in Figure 3b the upper left area (i.e., aggregated energy higher than $10^{-3}$ and distance from top nodes smaller than 3) is the most interesting one for a website administrator. Pages in that area have a reasonable probability to be visited while not being too far away from the home page. Website administrators can now utilize our methodology to test different biases and identify the one that best meets their requirements. For example, if the aim is to keep visitors of *DEM* close to the top pages but still enable them to easily explore other pages up to a distance of 2, the squared degree bias would exactly fulfill these requirements (cf. Figure 3b).

Please note that in all our calculations the values for the damping factors that we used apply for the general Web and may not hold for a particular website. However, for a given website the operators can determine the damping factors from the actual log-files.

**Findings & Implications.** Due to a popularity bias some nodes slide into a ground state in which they are almost never visited by the random surfer. The distance from top nodes determines the final energy of a node. Contrary, the unpopularity bias increases the flow of energy towards nodes with an initially very low energy. This means that an induced popularity bias increases the visibility of already frequently visited nodes and at the same time it shifts many other pages into the ground state. Pages in that state will be hardly visited. If the aim of a website is to be easily explorable (e.g. Wikipedia datasets *WFS* and *BW*) this should be taken into account. The same applies for web shops, such as *MS*, for which administrators might expect to increase sales of the top products by inducing a popularity bias. However, this will only increase the visits of popular pages—which we find to be mostly overview pages such as *games for* 5 *players* [9]—while putting many actual product pages into the ground state. Nevertheless, taking into account the expected session length of users, it can make sense to concentrate their attention on the top nodes, as they are unlikely to visit nodes further away from the home page.

## VIII. Conclusions and Future Work

In this paper we presented an approach for measuring the impact of and between biased random surfers and applied it

[9] http://www.milan-spiele.de/nach-anzahl-fuenf-spieler-c-93_98.html

to seven empirical datasets to highlight practical implications for different kinds of networks.

The results gathered from our experiments broaden our understanding of the impact of intrinsic biases for the random surfer on directed webgraphs. Additionally, we found that some combinations of measures and biases (e.g., penalization of popular pages decreases the probability of trapping the random surfer) perform consistently over all datasets. On the other hand, some results highly vary across experiments (e.g., the entropy rate of some biases depend on the structure of the network).

Regarding the *Website Coverage*, we conclude that all used biases highly influence visit probabilities of the random surfer. In particular, we find a consistent pattern based on the type of the bias: Popularity biases tend to trap the random surfer within just a few web pages of the website, whereas biases penalizing popular pages are able to increase website coverage.

The changes in *Surfer Guidance*, due to different biases, are more dependent on the network structure than on the type of the bias itself. However, all biases were able to decrease the entropy rate, which further indicates an increase in guidance.

For the *Web Page Response*, in networks with a large diameter we observed a strong side effect. Specifically, the bias puts many pages into a so-called ground state. Pages in that state are barely visited by the random surfer. Thus, website administrators should take these side effects into account.

For future work we plan on analyzing the influence of similarity-based as well as extrinsic biases on random surfers, such as text similarity or categorical mappings between articles (encyclopedias) or products (web shops). Further, we are interested in coloring nodes regarding their type (e.g., product pages, administration pages, types/categories of article pages) and analyzing which type of nodes are favored by different types of biases.

## ACKNOWLEDGMENT

This research was in part funded by the FWF Austrian Science Fund research project "Navigability of Decentralized Information Networks" (P 24866-N15).

## REFERENCES

[1] S. Al-Saffar and G. Heileman. Experimental bounds on the usefulness of personalized and topic-sensitive pagerank. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 671–675, Nov 2007.

[2] M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Trans. Internet Technol.*, 5(1):92–128, Feb. 2005.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[4] J.-C. Delvenne and A.-S. Libert. Centrality measures and thermodynamic formalism for complex networks. *Physical Review E*, 83(4):046117, 2011.

[5] L. Demetrius and T. Manke. Robustness and network evolution—an entropic principle. *Physica A: Statistical Mechanics and its Applications*, 346(3–4):682 – 696, 2005.

[6] A. Fronczak and P. Fronczak. Biased random walks in complex networks: The role of local navigation rules. *Physical Review E*, 80(1):016107, 2009.

[7] F. Geigl, D. Lamprecht, R. Hofmann-Wellenhof, S. Walk, M. Strohmaier, and D. Helic. Random surfers on a web encyclopedia. In *Proc. of the 15th International Conference on Knowledge Technologies and Data-driven Business*, i-KNOW '15, pages 5:1–5:8, New York, NY, USA, 2015. ACM.

[8] F. Geigl, K. Lerman, S. Walk, M. Strohmaier, and D. Helic. Assessing the navigational effects of click biases and link insertion on the web. In *Proc. of the 27th ACM Conference on Hypertext and Social Media*, HT '16, pages 37–47, New York, NY, USA, 2016. ACM.

[9] R. Ghosh and K. Lerman. Rethinking centrality: The role of dynamical processes in social network analysis. *Discrete and Continuous Dynamical Systems Series B*, 19(5):1355 – 1372, July 2014.

[10] D. F. Gleich, P. G. Constantine, A. D. Flaxman, and A. Gunawardana. Tracking the random surfer. In *Proc. of the 19th international conference on World wide web - WWW '10*, page 381, 2010.

[11] I. Goldhirsch and Y. Gefen. Biased random walk on networks. *Phys. Rev. A*, 35:1317–1327, Feb 1987.

[12] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 478–479, New York, NY, USA, 2004. ACM.

[13] D. Helic, M. Strohmaier, M. Granitzer, and R. Scherer. Models of human navigation in information networks based on decentralized search. In *Proc. of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 89–98, New York, NY, USA, 2013. ACM.

[14] N. Hill and D.-P. Häder. A biased random walk model for the trajectories of swimming micro-organisms. *Journal of Theoretical Biology*, 186(4):503–526, 1997.

[15] S. Hwang, D. S. Lee, and B. Kahng. Effective trapping of random walkers in complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(4), 2012.

[16] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[17] A. N. Langville and C. D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[18] K. Lerman and T. Hogg. Leveraging position bias to improve peer recommendation. *PLoS ONE*, 9(6):e98914, 06 2014.

[19] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[21] W. Parry. Intrinsic markov chains. *Transactions of the American Mathematical Society*, 112(1):pp. 55–66, 1964.

[22] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *Proc. of the 15th international conference on World Wide Web*, pages 727–736. ACM, 2006.

[23] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *NIPS*, pages 1441–1448, 2001.

[24] R. Sinatra, J. Gómez-Gardeñes, R. Lambiotte, V. Nicosia, and V. Latora. Maximal-entropy random walks in complex networks with limited information. *Physical Review E*, 83(3):030103, 2011.

[25] P. Singer, D. Helic, B. Taraghi, and M. Strohmaier. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PloS one*, 9(7):e102070, 2014.

[26] S. Walk, D. Helic, F. Geigl, and M. Strohmaier. Activity dynamics in collaboration networks. *ACM Transactions on the Web (TWEB)*, 10(2):11, 2016.

[27] S. Walk, P. Singer, L. E. Noboa, T. Tudorache, M. A. Musen, and M. Strohmaier. Understanding how users edit ontologies: comparing hypotheses about four real-world projects. In *International Semantic Web Conference*, pages 551–568. Springer, 2015.

[28] S. Walk, P. Singer, M. Strohmaier, T. Tudorache, M. A. Musen, and N. F. Noy. Discovering beaten paths in collaborative ontology-engineering projects using markov chains. *Journal of biomedical informatics*, 51:254–271, 2014.

[29] R. West and J. Leskovec. Automatic versus human navigation in information networks. In *ICWSM*, 2012.

[30] R. West and J. Leskovec. Human wayfinding in information networks. In *Proc. of the 21st international conference on World Wide Web*, pages 619–628. ACM, 2012.

[31] W. Woess. Random walks on infinite graphs and groups—a survey on selected topics. *Bulletin of the London Mathematical Society*, 26(1):1–60, 1994.