HopRank: How Semantic Structure Influences Teleportation in PageRank (A Case Study on BioPortal)

Lisette Espín-Noboa
GESIS & Uni. Koblenz-Landau
lisette.espin@gesis.org

Florian Lemmerich
RWTH Aachen University
florian.lemmerich@cssh.rwth-aachen.de

Simon Walk
Detego GmbH
s.walk@detego.com

Markus Strohmaier
RWTH Aachen University & GESIS
markus.strohmaier@cssh.rwth-aachen.de

Mark Musen
BMIR-Stanford
musen@stanford.edu

ABSTRACT

This paper introduces HopRank, an algorithm for modeling human navigation on semantic networks. HopRank leverages the assumption that users know or can see the whole structure of the network. Therefore, besides following links, they also follow nodes at *certain distances* (i.e., k-hop neighborhoods), and not at random as suggested by PageRank, which assumes only links are known or visible. We observe such preference towards k-hop neighborhoods on BioPortal, one of the leading repositories of biomedical ontologies on the Web. In general, users navigate within the vicinity of a concept. But they also "jump" to distant concepts less frequently. We fit our model on 11 ontologies using the transition matrix of clickstreams, and show that semantic structure can influence teleportation in PageRank. This suggests that users—to some extent—utilize knowledge about the underlying structure of ontologies, and leverage it to reach certain pieces of information. Our results help the development and improvement of user interfaces for ontology exploration.

CCS CONCEPTS

• **Information systems** → **Content ranking**; Browsers; • **Mathematics of computing** → *Exploratory data analysis*.

KEYWORDS

Biased random walker; PageRank; k-hop neighborhood; BioPortal

1 INTRODUCTION

Ontology Engineering and Ontology Learning are two branches of the Semantic Web whose aim is to accurately build and curate ontologies. The former studies new techniques to improve collaboration among humans while editing ontologies [26, 29], and the latter introduces new methodologies and algorithms to automatically create ontologies by crawling the Web [5, 25]. These efforts represent significant advances in the development of knowledge bases, which represent facts about the real world (e.g., people, diseases). However, there is little knowledge about how users consume such ontologies on the Web. To this end, Walk et al. studied how users browse BioPortal [28]. Their findings suggest that some ontologies influence the way users interact with the website. However, how users navigate through the ontology structure (i.e., from one concept to another) remains unclear.

Problem Statement: In this paper, we study the influence of semantic structure on *teleportation* (i.e., jumping to any node chosen at random) in PageRank. For example, consider the ontology shown in Figure 1(a), where nodes represent *classes* (a.k.a. concepts) and edges *isASubClassOf* relationships. On BioPortal, ontologies are shown vertically as hierarchical trees, and concepts can be explored using the *expand-on-demand* principle. This means that only top level concepts are shown first, and then users are able to expand and collapse as many concepts as they need at any level of the ontology. In other words, users can use and therefore are potentially aware of a *virtually fully connected network* in all stages of navigation. Previous studies [21, 31] have modeled user navigation using PageRank. However, these assume that navigation paths are constrained by links and random teleportation. In our scenario, where the whole structure of an ontology can be visualized at any time, we believe that teleportation is not fully random, but rather biased towards k-hop neighborhoods.

Approach: Motivated by previous studies on information foraging [8–10, 22], decentralized search [16, 18], and PageRank [4, 14, 15, 21, 31], we propose HopRank, a method for modeling transitions across k-hop neighborhoods on semantic networks. The key idea of this work relies on the HopPortation vector $\vec{\beta}$, which defines the probabilities of transitioning to each k-hop neighborhood. From the PageRank point of view, we can say that teleportation is not fully random, and the probability of following the structure of a page is not based only on one parameter (i.e., probability of following links), but on k parameters, representing all k-hop neighborhoods *reached from the current page*. Technically, we pass the HopPortation vector to a random walker to make biased decisions on which neighborhood to go next. Once this decision is made, the random walker uniformly chooses a concept within that neighborhood.

	β_0	β_1	β_2	β_3	β_4
Transitions	0	1	100	0	15
Smoothing	1	2	101	1	16
Normalized	0.008	0.017	0.835	0.008	0.132

	a	b	c	d	e	f	g
a	0.001	0.011	0.011	0.244	0.244	0.244	0.244
b	0.008	0.001	0.963	0.008	0.008	0.006	0.006
c	0.008	0.963	0.001	0.006	0.006	0.008	0.008
d	0.419	0.018	0.009	0.001	0.419	0.067	0.067
e	0.419	0.018	0.009	0.419	0.001	0.067	0.067
f	0.419	0.009	0.018	0.067	0.067	0.001	0.419
g	0.419	0.009	0.018	0.067	0.067	0.419	0.001

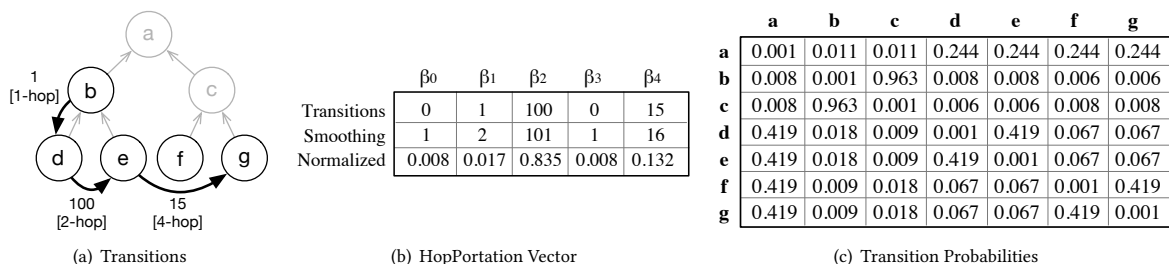(a) Transitions (b) HopPortation Vector (c) Transition Probabilities

Figure 1: HopRank on semantic networks. This example illustrates an instance of navigation on an ontology. (a) Shows the underlying network composed by seven concepts (a-g) and six *isASubClassOf* relationships (straight-thin grey arrows). Transitions (curved-thick black arrows) are labeled by the actual number of transitions between concepts, as well as the [k-hop] distance (i.e., shortest path) between them. (b) Illustrates how the HopPortation vector $\vec{\beta}$ is built using transition counts per k-hop. (c) Shows the transition probabilities inferred by HopRank, see Equation (1).

Contributions: The contributions of this paper are:

(1) We empirically show how users leverage the structure of the ontologies on BioPortal by quantifying the proportion of transitions per k-hop neighborhood.

(2) We propose HopRank, an algorithm for modeling human navigation on semantic networks.

(3) We demonstrate that HopRank outperforms traditional navigation and popularity-based models on BioPortal, especially when users browse ontologies directly without search.

(4) We make an implementation of this approach openly available on the Web [11].

2 RELATED WORK

BioPortal provides users with a *tree-like explorer* and a *local search* engine to navigate ontologies. In addition, concepts can be *expanded on demand* to see their children nodes. Although these functionalities are exploited differently across ontologies [28], it is unclear how users navigate through the ontology structure. Thus, this section covers previous work on search and navigation on networks.

Search. *Information Foraging* [22] assumes that people, when possible, modify their strategies or the structure of the environment to maximize their gain of valuable information. These patterns are also found in the way humans recall information from memory [17]. Similarly, *berrypicking* [6], a model of online searching, states that queries are not static, but rather evolve, and users commonly gather information in pieces instead of in one large set.

Navigation. *PageRank* [21] is the most popular method to measure the importance of web pages based on their incoming and outgoing links. It relies on an imaginary surfer who is randomly *clicking on links*, and eventually *jumps* to any node in the network. The probability of following links is given by a *damping factor*. Multiple variations have been proposed for improving information retrieval systems, e.g., a biased PageRank [15] to capture the importance of a page more accurately by taken topics into account or a weighted PageRank [31] to assign larger rank values to more popular pages (i.e., preferential attachment) instead of distributing the rank value of a page uniformly to all outgoing pages. Geigl et al. suggest that the behavior of a *random surfer* is almost similar to real users, as long as they do not use search engines [13]. They also find that

classical navigation structures, such as navigation hierarchies or breadcrumbs, only exercise limited influence on navigation. Experiments in [24] reveal that memory-less *Markov chains* represent a quite practical model for human navigation on a page level. However, this assumption is violated when the analysis is expanded to a topical level. Helic et al. identify certain configurations of *decentralized search* that are capable of modeling human navigation in information networks [16]. Their findings suggest that navigation on such networks is a two phase process combined with the *exploitation of the known* (i.e., goal-seeking) and the *exploration of the unknown* (i.e., orientation).

User Interfaces. Human navigation has also been studied for enhancing interfaces. For instance, [12] explores *fisheye* views to display large information structures such as programs and databases. The intuition behind this paradigm is that users often explore their neighborhood, and distant major landmarks in more detail. Similarly, Van Ham and Perer studied the *search, show context, expand on demand* browsing model in [27], and proposed techniques to design better graph visualization tools.

We propose HopRank—a biased random walker—to model navigation on semantic networks. HopRank builds upon insights from information foraging [17, 22], decentralized search [16, 18] and PageRank [21]. More precisely, we replace the *damping factor* by a HopPortation vector to encode the probabilities of visiting each k-hop neighborhood. The intuition here is that users browse semantically close terms more often than semantically distant ones.

3 BIOPORTAL

There exist a large number of ontologies in the biomedical domain. They are highly specialized and therefore expensive to develop. To enable ontology adoption and reuse, effective support for browsing and exploring existing ontologies is required. Towards that goal, the National Center for Biomedical Ontology (NCBO) [3, 19] features BioPortal [1, 20, 30]—one of the leading repositories of biomedical ontologies on the Web—containing currently more than 700 ontologies with more than 9 million ontology classes. On BioPortal, practitioners and experts can access ontologies via Web services and Web browsers. The latter allows users to navigate ontologies

by searching specific classes, or by directly browsing their concept hierarchies within a tree-like explorer [28].

Ontologies. We propose to model human navigation on semantic networks using the structure of the underlying ontology. On Bio-Portal, ontologies are defined as directed networks, where nodes represent *concepts* and edges *isASubClassOf* relationships. Since such edges are usually non-cyclic and have a common root, these ontologies often form trees. Table 1 shows 11 of the most visited ontologies in 2015[1]. For instance, LOINC the largest ontology with 175K nodes, 153K edges, and 74K connected components.

Transitions. We analyzed all HTTP requests made in 2015 and extracted 336K valid sessions (i.e., after filtering out sessions with less than 2 requests, and requests to ontologies or concepts which do not exist). Each session contains transitions (i.e., a sequence of visited concept pages) triggered by a single user (i.e., IP address) without breaks (i.e., pauses of at least 60 minutes). For simplicity, we only consider transitions within the largest connected component (LCC) of each ontology, and discard ontologies with less than 1000 transitions[2]. Overall, we found 11 ontologies and 133K transitions between their concepts[3], see Table 1 for some key properties.

Navigation Types. Based on the HTTP request headers, we inferred 7 navigation types: Details (DE), Direct Click (DC), Direct URL (DU), Expand (EX), External Link (EL), External Search (ES), and Local Search (LS). **DE**: are all clicks made within the *Details* tab of a selected concept. **DC**: are all clicks made on concepts within the tree-like explorer. **DU**: refers to all concept requests without HTTP referrer (e.g., direct URL in the browser). **EX**: considers all clicks on the (+) symbol of a concept, which triggers the expansion of the concept to show all its children nodes. Notice that this request is called only once, even if the symbol is clicked multiple times. The

---

[1]As ontologies can be edited over time, we work with their latest snapshots from 2015.
[2]Transitions within the LCCs of these ontologies represent 80% of all transitions.
[3]We left out the popular SNOMEDCT ontology due to computational limitations.

Table 1: Datasets. This table illustrates network properties of 11 of the most popular ontologies on BioPortal in 2015. Ontologies represent networks whose nodes refer to *concepts* and edges *isASubClassOf* relationships. Original number of nodes, edges, and connected components of ontologies are shown under N, E and cc, respectively. Properties of the largest connected component (LCC) of each ontology are shown under N', E', d' and T', where d' refers to the diameter and T' to the number of transitions.

| #	Ontology	N	E	cc	N'	E'	d'	T'
1	CPT	13219	13235	3	13092	13110	15	44651
2	MEDDRA	66506	31863	43493	22889	31738	8	42746
3	NDFRT	35019	34504	522	32074	32080	24	22452
4	LOINC	174513	152683	73518	100871	152558	13	6349
5	ICD9CM	22534	22531	3	22407	22406	12	4434
6	WHO-ART	1852	2997	3	1725	2872	4	2811
7	MESH	165166	24182	145652	16947	21596	31	2623
8	ICD10	12446	11256	1190	11132	11131	10	2288
9	CHMO	2966	3071	3	2964	3071	22	1423
10	HL7	10319	10600	1049	9146	10475	19	1374
11	OMIM	81821	39359	44110	37587	39234	6	1291
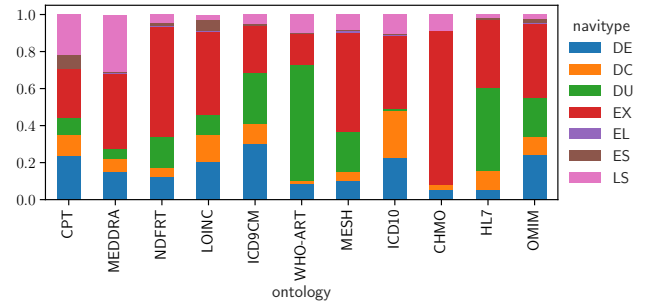
Figure 2: Navigation Types. Each bar shows the fraction of transitions within the LCC of each ontology. Stacked bars differentiate types of navigation: details (DE, blue), direct click (DC, orange), direct URL (DU, green), expand (EX, red), external link (EL, purple), external search (ES, brown) and local search (LS, pink). Most ontologies are mainly navigated by *expanding* nodes within the tree-like explorer.

opposite behavior (collapse) is not considered[4]. **EL**: captures all requests coming from external websites that are not search engines. **ES**: are all requests coming from the top 10 most popular external search engines such as Google and Yahoo. **LS**: are all requests made via the local search functionality of each ontology. Notice that this search is a 3-step process. First users type a keyword, then the system shows auto-suggestions and finally users click on one of the concepts shown in the auto-suggestion list. We only consider the final step a local search transition. **ALL**: includes all the above-mentioned types. Figure 2 shows the distribution of transitions across navigation types for each ontology. In general, most traffic comes from expanding a concept (EX, 44%), followed by local search (LS, 17%), direct URL (DU, 16%) and details (DE, 14%). Surprisingly, direct clicks on concepts (DC) only represent 7% of all transitions. This suggests that users spend substantial time expanding concepts before they find a concept of interest.

4 HOPRANK: A BIASED RANDOM WALKER

HopRank models human navigation on semantic networks. Imagine a random walker whose decisions on where to go next are biased towards specific k-hop neighborhoods. This bias is what we call *HopPortation*, which encodes the probabilities of transitioning to each k-hop neighborhood. In our model, navigation on networks can be explained as a 2-step process. First, a k-hop neighborhood of the current node i is drawn from a categorical distribution. Second, a node j is randomly chosen within that k-hop neighborhood. Note that this process holds only if the walker is fully or partially aware of the structure of the network (i.e., knows or can see it). Without this prerequisite, and if links are not preferred, then random jumps to random pages will be more plausible. In comparison to the classic random walker with teleportation (e.g., PageRank [21]), where its movements are constrained by the damping factor α (i.e., probability of following links), HopRank is constrained by a vector $\vec{\beta}$ containing

---

[4]Collapse is a client-side functionality, and thus, it is not recorded in the log files.

k different factors, which define the probabilities of going to each k-hop neighborhood from the current location.

Visited k-hop Neighborhoods on BioPortal. We aggregate ALL transitions by the shortest distance between two sequentially visited nodes. This distance is referred to as k-hop neighborhood. In Figure 3(a) we see that target nodes at large distances are less likely to be visited next. This is expected, since—to some extent—larger distances enclose more branches, therefore more target candidates. Note that ontologies are sorted by diameter in descendant order from MESH to WHO-ART. Interestingly, users tend to hop as far as the ontology's diameter, for $d' \leq 12$. For instance, OMIM's diameter is 6 (see Table 1), and 6 is the maximum hop done by users. Otherwise, users (roughly) hop up to two-thirds of the ontology's diameter, for $d' > 12$. For example, MESH's diameter is 31, and the largest hop reached is 19.

Transitions per k-hop Neighborhood on BioPortal. Figure 3(b) shows the average percentage of transitions across k-hop neighborhoods per navigation type. We see that users on average (ALL, grey) prefer to navigate through 2-hop (41%) and 1-hop (23%) neighbors. In particular, when navigation is triggered by direct clicks (DC, orange) and expand (EX, red). Notice their fast decay when $khop > 8$. Other types of navigation such as external link (EL, purple), and direct URL (DU, green)—which do not leverage the tree-like explorer—tend to reach concepts at larger distances more frequently. Notice their peaks at $khop = \{5, 11\}$, respectively. Interestingly, when users opt for external search (ES, brown), they often click on 2-hop concepts, but also on 12-hop and 15-hop neighbors. Intuitively, the details tab (DE, blue) helps users to click on nearby concepts at $khop \leq 2$, more often than local search (LS, pink), which is more likely to reach concepts at $khop \geq 2$.

5 MODELS OF HUMAN TRANSITIONS

In this section, we formally introduce our HopRank model, and recap popular navigation models for comparison. We denote the transition probabilities, and # of parameters according to HopRank and 7 other models that we will use later on for model selection.

We formally represent an ontology[5] as a graph $G = (V, E)$, with $V = (v_1, \ldots v_n)$ being a set of N nodes, and $E = \{(v_i, v_j)\} \in V \times V$ a set of undirected edges[6]. The ontology structure is captured by the adjacency matrix $A_{N \times N} = a_{ij}$, where a_{ij} is 1 if the link exists, 0 otherwise. Transitions are represented by the transition matrix $T_{N \times N} = t_{ij}$, where t_{ij} represents the number of transitions between source node i and target node j.

HopRank (HR). Given the HopPortation vector $\vec{\beta}$, the probability of reaching a k-hop neighborhood is denoted by factor $\beta_k \in \vec{\beta}$. The stochastic k-hop matrix M_k encodes all nodes j with a shortest distance k from i. HopRank uniformly distributes β_k across all nodes j at distance k. The limits of k-hop neighborhoods go from 1 (direct edges), to d', the diameter of the ontology G. Noise $\beta_0 = 1 - \sum_{k=1}^{d'} \beta_k$ is added to allow for random jumps and self-loops. Figure 1(b) illustrates how the HopPortation vector is computed

---

[5]We focus on its largest connected component (LCC)
[6]Directionality of edges is omitted to calculate shortest paths between all pair of nodes.

(a) % of dyads traversed per ontology
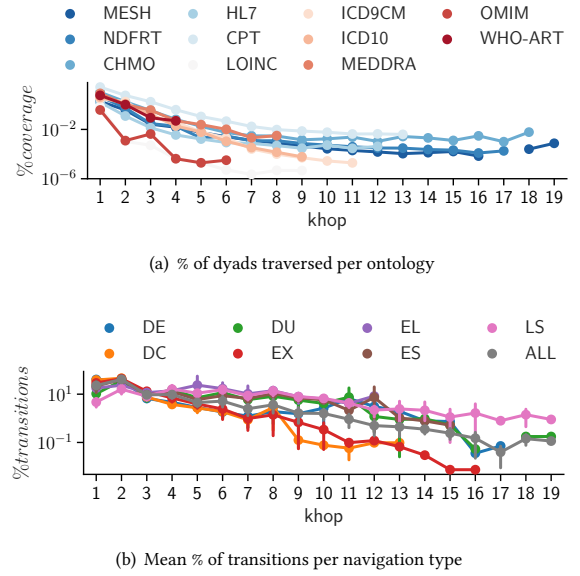

(b) Mean % of transitions per navigation type

Figure 3: Popularity of k-hops. (a) Shows the percentage of dyads that are traversed per k-hop neighborhood. Lines represent ontologies and are sorted by their LCC diameter: In descendent order from MESH (dark blue) to WHO-ART (dark red). (b) Shows the distribution of transitions across k-hop neighborhoods per navigation type. Percentages are averages across ontologies, and error bars the respective standard deviation. While several k-hop distances are being traversed non-uniformly, most transitions happen across nearby nodes, especially when browsing (DE, DC, EX, ES) 2-hop neighbors. In contrast, non-browsing types (EL, LS, DU) tend to reach more distant nodes more frequently.

from the transition counts. *Number of model parameters:* $d' + 1$.

$$P_{HR} = \beta_1 M_1 + \beta_2 M_2 + \cdots + \beta_k M_k + \frac{\beta_0}{N} \quad (1)$$

Preferential Attachment (PA). Given the degree matrix $D_{N \times N} = d_{ij} = d_j$, where d_j represents the degree of the target node j. The probability of moving from i to j is proportional to the degree of j. *Number of model parameters:* 0.

$$P_{PA} = D \quad (2)$$

Gravitational (Gr). Given the matrix $S_{N \times N} = (sp(i,j) + \epsilon)^2$, where $sp(i,j)$ denotes the shortest path between nodes i and j. The probability of navigating from i to j is proportional to the degree of node j and inversely proportional to the square distance between i and j. We add a smoothing factor ϵ to avoid overflows when dyads are disconnected. In such cases, we set ϵ to the diameter d' of G plus 1, to consider these jumps with a very low probability. Similarly, we set the diagonal (i.e., self-loops) to $\epsilon = d' + 2$. *Number of model parameters:* 0.

$$P_{Gr} = \frac{D}{S} \quad (3)$$

Random Walker (RW). Given the damping factor α (i.e., probability of following links), the probability of visiting a node j is proportional to α divided by the degree of the source node i, plus a random choice equally distributed among all nodes. Depending on the α value, a random walker can model four different behaviors: **(i)** $\alpha = 0.0$: random jumps only, **(ii)** $\alpha \approx 1.0$: navigation over links only, **(iii)** $\alpha = 0.85$: PageRank using the commonly used damping factor for navigating the Web [7], and **(iv)** the empirical PageRank which learns the parameter α from the transitions data. *Number of model parameters:* 1 *if empirical,* 0 *otherwise.*

$$P_{RW} = \alpha \boldsymbol{A} + \frac{(1 - \alpha)}{N} \qquad (4)$$

Markov Chain (MC). We assume that moving to the next node follows a Markov process. Therefore, the probability of moving to a node j only depends on the current node i. These probabilities represent the maximum likelihood, learned from the transition matrix T. Thus, the probability of visiting node j from node i is proportional to the number of transitions t_{ij}. *Number of model parameters:* $N \times (N - 2)$.

$$P_{MC} = \boldsymbol{T} \qquad (5)$$

Note that $\boldsymbol{M}$, $\boldsymbol{A}$, and all P_* from Equations (1) to (5) are right stochastic matrices (i.e., each row must sum to 1).

6 EXPERIMENTS

In this section, we compare the performance of HopRank to the baselines on synthetic and real-world networks.

6.1 Model Selection

For comparing the models, we employ the Bayesian Information Criterion (BIC) [23] to select the best, i.e., lowest BIC score. BIC evaluates *log-likelihoods LL* (i.e., how likely our transitions are for a given model) and takes into account the *number of model parameters* and *observations* (i.e., # of transitions) to avoid over-fitting.

$$BIC = -2 \cdot LL + nparams \cdot log(nobservations), \qquad (6)$$

$$LL = \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} \cdot log(p_{ij}), \qquad (7)$$

where t_{ij} represents the actual number of transitions from node i to node j, and p_{ij} the probability of transitioning from node i to node j for a given model.

6.2 Synthetic Network

Setup. The underlying network (structure) is a binary tree composed by $N = 7$ nodes and $|E| = 6$ edges as shown in Figure 1(a). Transitions (curved-thick edges) are biased towards 2-hop and 4-hop neighborhoods. These biases are reflected in the HopPortation vector shown in Figure 1(b).

Results. Probabilities inferred using Equation (1) are depicted in Figure 1(c). Figure 4 (left) shows the number of parameters inferred by each model. While the Markov chain model (MC) requires 35 parameters, HopRank only needs 5. The empirical PageRank (RW E.) learned a damping factor of $\alpha = 0.01$. This means that users are 1% likely to follow links. In Figure 4 (right) we see the comparison of models using BIC scores. In this synthetic network, transitions
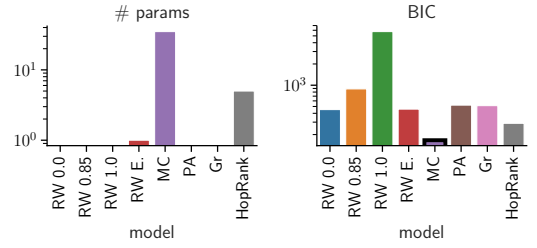

Figure 4: Results on Synthetic Network from Figure 1. X-axis maps the models at interest. (Left) Number of parameters inferred by each model. (Right) BIC scores: the lower the better at explaining the data. In this example, navigation is best described by Markov chain followed by HopRank.

are best described by the Markov chain model because model parameters (i.e., maximum likelihood) are proportional to the actual transition counts per dyad, and the data structure is very small[7]. In spite of that, HopRank is the second best model and describes navigation better than random (RW 0.0).

6.3 Medical Dictionary for Regulatory Activities Terminology (MEDDRA)

Setup. MEDDRA[2] is one of the the largest ontologies in our dataset (see Table 1). After pre-processing, its largest connected component (LCC) consists of $23K$ nodes and $43K$ transitions.

Results. Figure 5(a) shows the HopPortation vectors learned for each type of navigation in MEDDRA. We see that users mainly navigate through 1, 2, 6, and 8-hop neighbors. For instance, transitions through direct clicks—on a concept (DC), its details (DE) or expand (EX)—mainly follow 1-hop and 2-hop neighbors. However, when transitions are triggered by direct URLs (DU), local search (LS) or external links (EL), users tend to reach distant target nodes (i.e., 6-hop and 8-hop neighbors). Figure 5(b) shows the ranking of models according to BIC scores (lower is better). We see that in MEDDRA all types of navigation are best explained by HopRank.

6.4 Top11 Ontologies in BioPortal

Setup. We fit HopRank and the baseline models to all transitions by ontology and navigation type. These represent $133K$ transitions coming from the 11 ontologies described in Table 1.

Results. In Figure 6 we highlight the model that explains the number of transitions per ontology and navigation type best (i.e., the model with lowest BIC score). Ontologies are sorted by their number of transitions from CPT (largest) to OMIM (smallest). HopRank outperforms the other models 89% of the time, especially when users browse directly—regardless of the ontology—the tree-like explorer via clicks (DC), details (DE) and expand (EX). When there are not enough observations (i.e., the number of transitions is small), the other models tend to outperform HopRank due to the fact that the other models require fewer parameters and/or it is less likely to find transitions across different k-hop neighborhoods. This is the case for 6 ontologies in certain navigation types. For instance, we found

---

[7]Therefore, number of parameters does not play a very important role in BIC.

(a) HopPortation vectors
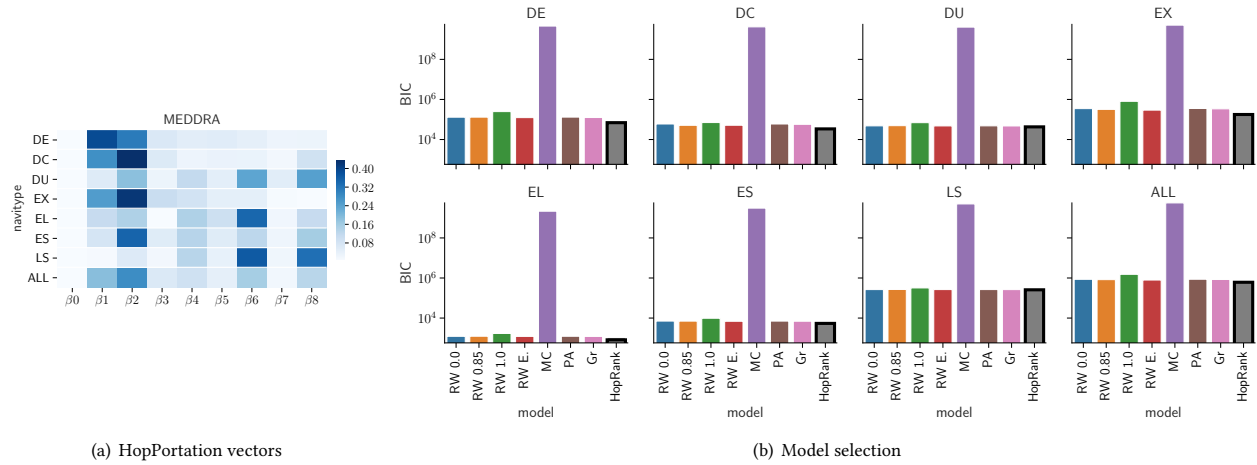
(b) Model selection

Figure 5: Results on MEDDRA. (a) This heatmap shows the HopPortation vectors learned from the transitions in MEDDRA. Cells represent the probabilities of visiting a certain k-hop neighborhood (column) by a given navigation type (row). In general, 2-hop and 1-hop neighborhoods are more likely to be visited next, regardless of navigation type (ALL). However, distant hops are preferred through *direct URLs* (DU), *external links* (EL), and *local search* (LS). (b) This figure shows the comparison of models across navigation types using BIC scores. We see that HopRank outperforms all baseline models.
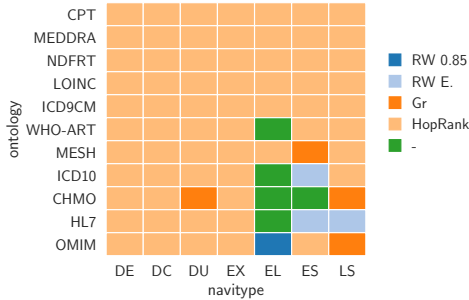

Figure 6: Model Selection on BioPortal. This heatmap highlights the model—with lowest BIC score—that best describes the # of transitions per ontology and navigation type. HopRank outperforms the other models 89% of the time, especially when browsing concepts via details (DE), direct click (DC) and expand (EX). When transitions are scarce (i.e., the other 11%), BIC penalizes HopRank since it has more parameters than the other models (except Markov chain).

5 external search (ES) transitions in MESH which are best described by the Gravitational model (Gr). Even though HopRank was a better candidate (i.e., higher log-likelihood), BIC penalized it for having more parameters ($nparams_{HopRank} = 32 > nparams_{Gr} = 0$). Notice that we model navigation in ontologies with at least 2 transitions. Ontologies that do not fulfil this condition per navigation type are marked as green cells "-".

7 DISCUSSION AND FUTURE WORK

In this section, we discuss decisions made for data processing, and future directions that can be pursued to improve our results.

Largest Connected Component (LCC). Surprisingly, ontologies on BioPortal may have multiple connected components. In those cases, only the branch connected to the root *owl:Thing* is shown at first in the tree-like explorer. Disconnected (and hidden) nodes or branches need to be accessed from external pages or local search. For simplicity, we opted to work with the LCC of each ontology with the cost of removing 20% of all transitions. Future work should consider the whole network to study the tradeoffs between number of transitions and random teleportation.

HopRank Extensions. More extensions based on network properties or similarity measures between nodes could improve our results. For instance, considering ontologies as directed graphs, and assuming that navigation is not only constrained by distance but also directionality: top-down or bottom-up.

Other Types of Networks. Even though this paper targets semantic networks, we believe that HopRank can be utilized to model human navigation in other networks, such as the Web or cities. The only assumption required is that users must have background knowledge of the underlying network they are surfing/traveling in.

8 CONCLUSIONS

In this paper, we introduced the concept of *HopPortation* which states that users—navigating a known or visible network—are biased towards certain k-hop neighborhoods. This is a variation of PageRank, where we assume that teleportation is not fully random but rather distributed non-uniformly across different neighborhoods. We proposed *HopRank*—a biased random walker—to model navigation on semantic networks. Our findings on BioPortal suggest that semantic structure (i.e., shortest path) influences navigation on networks. In particular, users tend to be biased towards certain k-hop neighborhoods depending on the type of navigation. For instance, when manually browsing the tree-like explorer, users tend to hop to nearby concepts, whereas far-away concepts are more

likely to be reached by non-browsing types such as external links. These results advance our understanding of how ontologies are actually navigated and consumed, and help to develop and improve user interfaces for ontology exploration.

REFERENCES

[1] 2011. BioPortal. https://bioportal.bioontology.org/ Accessed: 2019-02-21.

[2] 2011. MEDDRA. https://bioportal.bioontology.org/ontologies/MEDDRA Accessed: 2019-02-21.

[3] 2011. National Center for Biomedical Ontology (NCBO). https://www.bioontology.org/ Accessed: 2019-02-21.

[4] Joshua T Abbott, Joseph L Austerweil, and Thomas L Griffiths. 2015. Random walks on semantic networks can resemble optimal foraging. (2015).

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. *The semantic web* (2007), 722–735.

[6] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.

[7] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30, 1-7 (1998), 107–117.

[8] Stuart K Card, Peter Pirolli, Mija Van Der Wege, Julie B Morrison, Robert W Reeder, Pamela K Schraedley, and Jenea Boshart. 2001. Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 498–505.

[9] Ed H Chi, Peter Pirolli, Kim Chen, and James Pitkow. 2001. Using information scent to model user information needs and actions and the Web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 490–497.

[10] Ed H Chi, Peter Pirolli, and James Pitkow. 2000. The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 161–168.

[11] Lisette Espín-Noboa. 2018. HopRank. https://github.com/lisette-espin/HopRank Accessed: 2019-01-23.

[12] George W Furnas. 1986. *Generalized fisheye views*. Vol. 17. ACM.

[13] Florian Geigl, Daniel Lamprecht, Rainer Hofmann-Wellenhof, Simon Walk, Markus Strohmaier, and Denis Helic. 2015. Random surfers on a web encyclopedia. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*. ACM, 5.

[14] Michael Gorman. 2004. Google and God's mind. *Los Angeles Times* 17 (2004).

[15] Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering* 15, 4 (2003), 784–796.

[16] Denis Helic, Markus Strohmaier, Michael Granitzer, and Reinhold Scherer. 2013. Models of human navigation in information networks based on decentralized search. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, 89–98.

[17] Thomas T Hills, Michael N Jones, and Peter M Todd. 2012. Optimal foraging in semantic memory. *Psychological review* 119, 2 (2012), 431.

[18] Jon M Kleinberg. 2002. Small-world phenomena and the dynamics of information. In *Advances in neural information processing systems*. 431–438.

[19] Mark A Musen, Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Christopher G Chute, Margaret-Anne Story, Barry Smith, and NCBO team. 2011. The national center for biomedical ontology. *Journal of the American Medical Informatics Association* 19, 2 (2011), 190–195.

[20] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research* (2009), gkp440.

[21] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab.

[22] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[23] Gideon Schwarz et al. 1978. Estimating the dimension of a model. *The annals of statistics* 6, 2 (1978), 461–464.

[24] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. 2014. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PloS one* 9, 7 (2014), e102070.

[25] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 697–706.

[26] Tania Tudorache, Jennifer Vendetti, and Natalya Fridman Noy. 2008. Web-Protege: A Lightweight OWL Ontology Editor for the Web.. In *OWLED*, Vol. 432.

[27] Frank Van Ham and Adam Perer. 2009. "Search, show context, expand on demand": supporting large graph exploration with degree-of-interest. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009).

[28] Simon Walk, Lisette Espín-Noboa, Denis Helic, Markus Strohmaier, and Mark A Musen. 2017. How Users Explore Ontologies on the Web: A Study of NCBO's BioPortal Usage Logs. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 775–784.

[29] Simon Walk, Philipp Singer, Lisette Espín-Noboa, Tania Tudorache, Mark A Musen, and Markus Strohmaier. 2015. Understanding how users edit ontologies: Comparing hypotheses about four real-world projects. In *International Semantic Web Conference*. Springer, 551–568.

[30] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids research* 39, suppl_2 (2011), W541–W545.

[31] Wenpu Xing and Ali Ghorbani. 2004. Weighted pagerank algorithm. In *Proceedings of the Second Annual Conference on Communication Networks and Services Research, 2004*. IEEE, 305–314.