Concept Hierarchies and Human Navigation

Salvador Aguinaga*, Aditya Nambiar[†], Zuozhu Liu[‡] and Tim Weninger*
*University of Notre Dame, Email: {saguinag, tweninge}@nd.edu,
[†]IIT Bombay, Email: aditya.nambiar@iitb.ac.in,
[‡]Zhejiang University, Email: lcowen.hn@zju.edu.cn

***Abstract*—We are confronted with massive amounts of information at every turn. In order to efficiently reason about knowledge and information, humans have evolved efficient strategies for organizing complex concepts in order to form connections between and recall information. This behavior can be observed and codified when people search for objects within digital information networks. Current models of search behavior exhibit unnecessary or extraneous complexity. Minimal or simple modifications to well established algorithms yield valid models of human navigation by exploring hierarchical information inherent in networks.**

We explore and validate a new model of how humans navigate an information networks. To that end, we present a new path finding algorithm that approximates human navigation by leveraging the categorical classification of the nodes within the network. We compare our new model, *CatPath*, to existing graph distance measures when possible and show that the category paths are largely correlated with traces of human navigation.

***Keywords*-navigation; information networks; Wikipedia**

I. Introduction

Large amounts of information are created and consumed by humans and machines at an unprecedented rate day after day. These data are maintained in a variety of information systems, which are responsible for organizing and retrieving the information when requested by a user. Research in information network analysis has shown that the organization of information plays a pivotal role in many search and retrieval tasks. Information *networks* systems organize data in ways that allows us to model data as a *graph* where graph-nodes can be related to other nodes in complex ways. The Web is a classic example of such an information network where Web pages are organized according to their hyperlinks. Humans can universally and naturally search for information on the Web by navigating paths of Web pages via their hyperlinks without instruction. Although human navigation on the Web has certainly changed since the wide adoption of search engines, these tools achieve their effectiveness by modeling how a human might navigate the Web by following hyperlinks.

However, many of these tools are poor representations of how humans actually navigate information networks. An example of a navigation model is the random surfer or random walker introduced to the field in the PageRank algorithm [1]. The PageRank surfer does not surf with a target in mind like most humans do. Alternative models have been developed to direct or otherwise un-randomize the walker, but these stochastic processes are typically used to solve some downstream task such as clustering, classification or retrieval [2].

Recent work has focused on modeling how humans choose these *walks*, henceforth known as *paths*. Humans possess access to abstract concepts to draw from when searching for information. We believe that connections between physical and abstract concepts lead to choices in path-finding. Many have studied path length [3], [4], [5] in part motivated by Milgram's seminal work on the *"The Small World Problem"* [6]. A study by West and Leskovec explores how people navigate information networks and solve wayfinding tasks [7]. Their key findings show that they can predict the information (target or destination page) the seeker is looking for from a short prefix of the navigation path.

More recently, work on human navigation of information networks looks at modeling the path by utilizing higher order Markov chain models, *i.e.*, memory models, as oppose to classic first order Markov chains, *i.e.*, memoryless models [8]. Models of navigation paths typically emphasize predicting the next node from a path prefix. Both bodies of work aim to elucidate human navigation at a more fundamental level. Probing the structure of information systems, observing, and capturing human behavior helps us build and augment.

Stemming from this mode of thinking, human navigation models in information networks were originally concerned with node similarity [9], where the human navigator always chooses the node with the smallest distance to the *destination* node [10], [7], [11], [12]. In these models, the destination node is known and the topological distance or term-document distance is applied to calculate which node is closest to the destination; and although this search strategy may indeed result in short or closer-to-optimal paths they may not actually mirror human behavior. Building on the work of those listed, our interest, more explicitly, is on the following questions: firstly, can conceptually shorter paths on local choices for navigation offer a more accurate model of human navigation in information networks?, and secondly, are the paths themselves more or less similar?

This article focuses on a new model of how humans search for a target in massive information networks. We approach

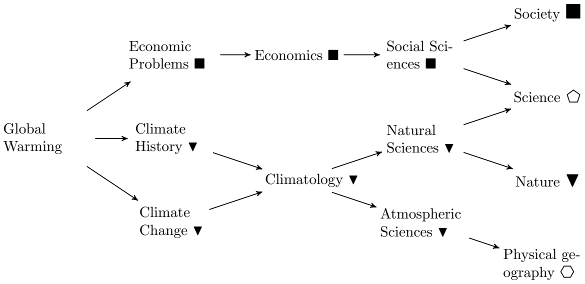


Figure 1: Type hierarchy on Wikipedia for Global warming. The symbols denote membership to a top-level category, *e.g.*, squares belong to the Society top-level category in Wikipedia, etc.

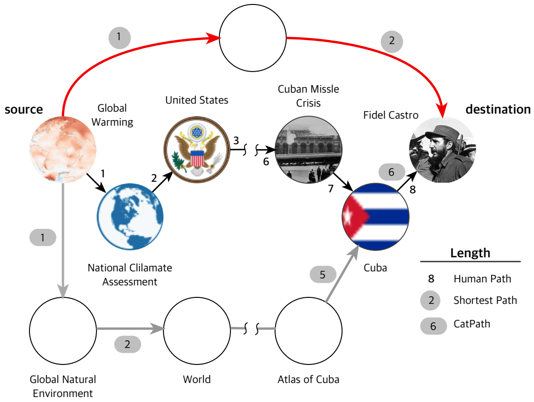


Figure 2: A representative example of a human path (HP) between concepts Global Warming and Fidel Castro. Nodes represent Wikipedia articles and the edges are hyperlinks clicked by the user (▬), Category Path (▬), and Shortest Path (▬)

this problem by first observing how the humans navigate information networks that contain some added descriptive schema or concept space. Our model leverages certain properties of the concept-space to describe the conceptual distance between two connected nodes in order to model patterns of human navigation. In contrast to the previous work, which makes navigation decisions based on distance to the final destination (*i.e.*, a global distance measure), our model explores both local *and* global distances when making navigation decisions. A natural example of this idea is Wikipedia, in which Wiki-articles and inter-article links compose an information network, and category pages exist as an overarching conceptual organization of the underlying Wiki-articles.

The elements along a search chain or path, from s to t (*i.e.*, source to target), contain associations humans develop from every day experience. These associations are used efficiently by people of varying age, background, etc. to find information or gain knowledge. These associations are analogous to a hierarchy where a concept is connected to a class or category of a higher level. Some examples of search chains that map nodes to higher-level types of concepts has been discussed in recent literature [13], [7], [8]. In the case of Wikipedia, Wiki-articles can be described by their membership in one or more Wikipedia categories. We are interested in using these categories, as presented in the structure of Wikipedia, to model how humans navigate Wiki-pages. An example of such a category hierarchy is shown in Figure 1, where Global warming is a member of three categories, Economic problems, Climate history and Climate change (among others not shown); these low-level categories are further members of middle-level categories, which are members of progressively higher level categories until one of the 22 main topic classifications, *e.g.*, Society, Science, Nature, is reached. In general, this work makes the following contributions: *1.)* We present a new model of human navigation paths, CatPath, that uses the conceptual category graph on Wikipedia as a model of how humans navigate information networks. *2.)* We validate the new model by comparing traces of human paths to CatPath as well as other path finding algorithms such as shortest path and personalized PageRank.

II. Category Paths

We model human navigation paths by computing Djikstra's shortest path algorithm on a weighted information network, where edge weights are computed to correspond to a shortest path through some concept space. For example, consider the illustration in Figure 2, which represents the Global warming node belonging to both, the article-network whose content links to other articles in a directed manner and belongs to a concept space also. Recall from Figure 2 that the shortest path between the source and destination nodes is actually only 2. Yet human navigators are typically unaware of the optimal paths between disparate nodes. So instead, humans are likely to pick a topologically non-optimal route that is deemed, by the walker, to move conceptually closer to the target [7]. Because Global warming is conceptually very distant from Fidel Castro intermediate steps need to be taken when navigating from one to the other. The CatPath model approximates human behavior by weighting each edge in the information network by its distance through the concept space.

Formally, we define an information network as $\{V, E\} = G$, where V is the set of all vertices and E is the set of all edges that connect two vertices. Many times, an information network has a concept space, schema or type system. In the CatPath model the vertices $v \in V$ are defined as being either concept-nodes (c-nodes, $*_c$) or physical-nodes (p-nodes, $*_p$) according to their function in the graph. The edges $e \in E$

are defined as being either a concept-edge (c-edge, $u_* \leftrightarrow v_c$) that connects any node to a c-node, or a physical edge (p-edge, $u_p \rightarrow v_p$) that connects two p-nodes. The delineation of concept versus physical is a problem and graph-dependent, but generally p-nodes correspond to concrete, atomic objects and c-nodes correspond to an amalgamation or classification of one or more other nodes.

Wikipedia is one example of such an information network, where Wiki-articles correspond to p-nodes, category pages correspond to c-nodes, Wiki-links are directed edges that connect two p-nodes, and category links are undirected edges that connect two c-nodes or a p-node with a c-node. In this example, Wiki-articles represent well defined, actual objects or ideas, whereas category nodes are conceptual classifications or agglomerations of Wiki-articles as well as other category nodes.

For a given source p-node $u_p \in V$ and destination p-node $v_p \in V$, a path through the information network is a sequence of edges that connects u_p to v_p denoted $u_p \rightsquigarrow v_p$. The shortest path between two nodes u_p and v_p is the shortest sequence of edges that connect u_p to v_p; ties are broken arbitrarily. Because of the differences between p-nodes and c-nodes defined in the CatPath model, we define two types of paths: physical paths (p-paths) and concept paths (c-paths aka *CatPaths*). A physical path exists between two p-nodes and must only contain other p-nodes $u_p \rightarrow x_p \rightarrow \cdots_p \rightarrow v_p$. A concept path also exists between two p-nodes but must only contain c-edges, *i.e.*, $u_p \rightarrow x_c \rightarrow \cdots_c \rightarrow v_p$.

The CatPath approach to human navigation assumes that humans navigate to nodes that are most conceptually related to their current node with respect to their destination. CatPath therefore re-weights p-edges incident to the current node according to length of the corresponding shortest c-edge. Formally, given a current node x_p with outgoing edges to node y_p and z_p, CatPath sets the weight of each edge to be the length of the corresponding shortest c-path between $x_p \rightsquigarrow_c y_p$ and $x_p \rightsquigarrow_c z_p$ respectively.

In the running example, the link between Global warming and World is given a weight of 7 because the shortest c-path through the category graph traverses 7 c-edges (5 category-to-category-edges, 2 category-to-article edges, and incidentally includes 6 category-nodes). Similarly the weight between Global warming and Global natural environment is 2, and the weight between Global natural environment and World is 2. With these weights computed a decision can be made about which node should be traveled to next.

In the general case, CatPath models conceptual separation of otherwise connected nodes in an information network by finding the shortest path through its concept-space. Adjacent p-nodes that are conceptually close together will have a short corresponding c-path and therefore be given low edge weights; conversely, adjacent p-nodes that are conceptually far apart are likely to have longer c-paths and therefore are likely to be given higher weights.

In order to efficiently compute the CatPath distance, each edge weight is computed lazily, that is, we take the category distance between two nodes x_p and y_p when performing a relaxation. This results in a nested-shortest path search, wherein the outer-shortest path search uses Dijkstra's algorithm to perform relaxations on the information network using weights computed via an inner-shortest path search. For our purposes Dijkstra's algorithm is tuned so that lower weights are better. The inner-shortest path search performs the Dijkstra-like breadth first search (BFS) on the unweighted concept network between the adjacent nodes from the outer-shortest path search. The length between these two adjacent nodes is returned to the outer-shortest path searcher and set as the weight between current x_p and a candidate next-node y_p.

Summary: CatPath models human path finding by performing shortest path search on a network where the edges are weighted by their distance in the concept hierarchy (*e.g.*Wikipedia's category hierarchy). We consider CatPath to be a local-global hybrid model because the edge-weights are computed locally, while the shortest CatPath from source to destination is inherently a global metric. In contrast, we view the human navigation model of Trattner *et al.* [11] (based on the social networks model of Adamic *et al.* [10]) as a global model because it greedily navigates to the adjacent node that is globally closest to the destination. Similarly, we view the ϵ-greedy model of Helic *et al.* as a global model because it too chooses to navigate to the adjacent node that is globally closet to the destination (albeit with a $1-\epsilon$ chance randomness [12]).

A. Implementation Details

The CatPath model must run on very large, complex datasets. Procedural programming methods, including Dijkstra's dynamic programming approach to shortest path finding, is infeasible for even moderately sized graphs. This is especially true for the nested path finding algorithm employed by CatPath. To address these issues we implemented the nested shortest path algorithm in the vertex programming paradigm using the Pregel-like framework, PowerGraph [14]. CatPath was implemented as a rather straightforward adaptation of the existing single source shortest path algorithm for vertex programming. The CatPath vertex program was able to compute the category distance from a given starting point to all other nodes in the 10 million node Wikipedia graph in a couple of hours on average.

III. Experiments

In order to determine the ability of CatPath to model human path-finding behavior, we compare paths of human navigation to the CatPath model as well as the shortest path and the personalized PageRank score [2] (*i.e.*, random walk with restart) on a large information network.

A. Datasets

The Wikipedia dataset used in this work was retrieved from the public data dump in December 2013. Wikipedia is an ideal example of an information network with a built in concept space. It consists of 10,276,554 pages, of which about 4 million are standard Wiki-articles, over 1 million category nodes, and 740 million total edges, other pages are redirects or disambiguation pages. In addition, we use a collection of human paths collected from the managers of the Wikipedia Game[1], an online and mobile game that places users at random Wikipedia pages and asks them to find another random Wiki-page. The Wiki Game setup is a nearly ideal human signal for our task for two reasons: 1) humans are given a destination and asked to find it in the fewest number of clicks possible, and 2) the Wiki Game is an untimed competition which allows human navigators to think carefully about their selections while providing a competitive incentive to perform as effectively as possible. The Wiki Game dataset was collected in June of 2013 and contains 1,966,704 games with 54,996 users. Within this set of games, 7,052 distinct source nodes and 3,497 distinct destination nodes are observed, see [15] for more details on these user-generated paths.

Note that the Wiki Game dataset uses the full Wikipedia network and is *different* from the Wikispeedia game. While the source and destinations in the Wiki Game appear artificial, keep in mind that two disparate pages make it difficult rely on intuition to easily make a decision on what link to follow. Disparate pages forces players to integrate knowledge of the two concepts and the workings of Wikipedia hyperlinks. We are, after all, interested in how the underlying network structure is leveraged in these tasks. The size of the combined dataset is significant in terms of computation and the memory footprint required to find answers to questions in a reasonable amount of time and with out excessive hardware needs. Moreover, the massive trails of clicks left behind by Wiki Game users, or from similar datasets, offer unprecedented opportunities to enhance our understanding of human behavior.

B. Methodology

Recall that the overarching goal of this paper is to develop a model that simulates HP finding in large information networks. With that in mind, we need to compare the proposed CatPath model to as many traces of human navigation as possible. A comparison methodology, therefore, ought to satisfy two objectives: (1) the test set should be large and diverse, and (2) because human navigation traces may be wildly inconsistent, the test set should contain as many traces of human paths as possible for each source and destination.

With these goals in mind, we selected the 100 most frequent source nodes from the 7,052 Wiki Game starting points, that is, we selected the 100 nodes that serve as the starting points for the most games. Using these 100 most frequent starting points, we found the CatPath distance, shortest path distance and personalized PageRank (PPR) scores for each possible target node. For example, the most common starting point was World War I, so CatPath, shortest path and PPR scores were generated between World War I and all other Wiki-articles. This process was repeated for the remaining 99 frequent starting points.

Alternative path sampling strategies are possible, but the above methodology does satisfy the stated objectives resulting in a test set that is large and contains many comparable human paths. In the remainder of this section we determine which path generation technique best approximates traces of human navigation.

1) Path Comparison Metrics: The working dataset is briefly described using the metrics shown in Table I after filtering out unsuccessful human paths (*i.e.*, Wiki Games that did not reach the destination), and human paths that exceeded 30 clicks (*i.e.*the extreme tail of the path length distribution).

Table I: Summary statistics of the dataset and KS-test statistics comparing the shape of path length distributions. Lower is better. * indicates p-value < 0.001. Kolmogorov-Smirnov goodness of fit tests (KS-test) were performed to determine how the shape and size of the distributions compare.

Path Metric	**Summary**				**KS-test**			
	ν	μ	**Mo**	σ	HP	CatPath	PPR	SP
Human Path	5	6.13	6	2.52	0	0.06*	1.0*	0.95*
CatPath	5	5.82	5	2.01		0	1.0*	0.98*
PPR	-	-	-	-			0	0.99*
Shortest Path	3	2.87	3	-				0

Contrary to the reports by West and Leskovec, who found that the median path lengths of human paths and shortest paths only differ by a length of 1 [7], we find that human paths are much longer than the mean and median as compared to the actual shortest paths. It is unclear why these results are so drastically different than the result of West and Leskovec; similar path filtering is performed in both cases and the sample sizes are both sufficiently large so as to preclude statistical anomaly. One possible explanation stems from the datasets used. The network used by West and Leskovec is a rather small educational subset of Wikipedia containing only 4,604, which is nearly 3 orders of magnitude smaller than the data set used in this experiment. It may be reasonable to assume that larger networks offer a greater opportunity for a user to veer off course and take longer paths on average. Another possible explanation for this discrepancy may be due to the differences between

[1] http://thewikigame.com/

the Wikispeedia dataset, developed and used in West and Leskovec's experiments, and the Wiki Game dataset used in this work. We will not enumerate the differences here except to say that it is possible, although we argue not particularly likely, that minute differences in gameplay may render comparison moot.

The results from Table I show a small sample of some positive results. The CatPath length and Human Path length (HPL) mean, median, mode and standard deviations seem, at first glance, to match quite well. The next section performs a through comparison of the proposed CatPath Model with Personal PageRank (PPR) [2], the shortest path and the recorded human paths. Related work is generally incomparable because, for example, West and Leskovec do not propose a model for human paths per se; instead, their work describes the form of human traces and uses those features to predict a user's destination [7]. Similarly, Adamic and Adar's work does not investigate traces human navigation, but rather topological separation due to email correspondence [10]. The work by Trattner *et al.* and the study by Helic *et al.* are the most relevant, but still they mainly focus on hierarchy induction and path similarity. The former used both, artificially induced hierarchies from external knowledge, an approach rooted in the well cited work of Benz *et al.* [16], and Wikipedias inherent category labels [11], [12]. Trattner *et al.* concluded that the inherent hierarchy of the network can approximate human navigation better than induced hierarchies from external knowledge. Thus, we limit our work to Wikipedia's category labels to build our concept space.

To compare the applicable models we look at the shape and content of the various paths with the following metrics:

· **Path Length** is the number of edges traversed in order to navigate from the starting point to the destination. This is indeed a simple metric, yet two paths of differing lengths are naturally not the same, and the dissimilarity between the two paths is naively proportional to the path length difference.

· **Path Distance** is the summation of the edge weights traversed as the user navigates from a source to the destination. Distinguishing path distance from path length is important in the context of this paper because edge weights are determined by the Category graph distance, which is critical to the underlying model. To compare path length and path distance we employ standard correlation statistics: Pearson's r, which measures the correlation between pairs of *values* in two lists, and Kendall's τ, which measure the correlation between pairs of *ranks* in two lists. A high correlation would indicate that the compared models are similar.

· **Distance-to-go** In addition to the total path length we also compare paths by observing how they change as the human navigator moves closer to the target. In this metric, we compare all possible states of navigation (at the beginning, with $l-1$ edges remaining, $l-2$ edges remaining, etc) in each applicable model. Agreement of two or more models in this metric would indicate similarity; disagreement in this metric would indicate the opposite.

· **Length-to-go** Just as path distance is similar to path distance as described above; length-to-go is analogously similar to distance-to-go.

· **Mean Difference** In addition to correlations, differences in path lengths is another indicator of path similarity. The smaller the mean difference the more similar the paths. This is similar to the *stretch* metric used in works by [11], [12].

· **Jaccard Coefficient** The previous methods have mostly focused on topological path measures. However, the content or labels of nodes can provide further insight into the similarity of the paths generated by various models. The Jaccard coefficient, defined as $(P \cap Q)/(P \cup U)$ for two sets P and Q, is the natural way to express similarity in terms of node/object overlap.

Taken together, these metrics provide a reasonable assessment toolbox that is used to identify the relative effectiveness of CatPath in the next section.

C. Results

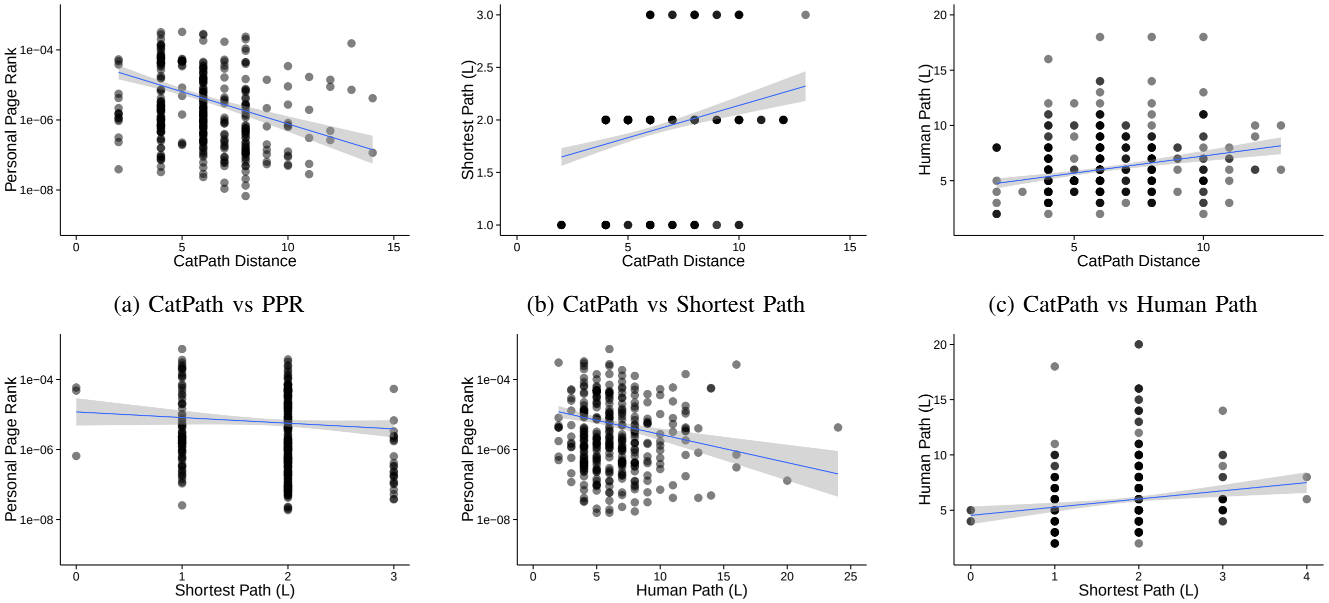
For each of the top 100 starting points (u) in the Wiki Game we computed the number of clicks needed to reach every end-point (v). The corresponding sum of weights in the resulting CatPath (*i.e.,* the CatPath distance) was computed for each $u \rightsquigarrow v$, as were shortest path length (SPL) and PPR scores. Pearson's correlation and Kendall's τ were computed on the resulting measurements and are listed in Table II.

Table II: Kendall's rank correlation τ and Pearson's product-moment correlation

	τ	τ p-value	r^2	p-value
CatPath – PPR	-0.165	$\leq$ 2.22e-16	0.0006	5.7e-06
CatPath – SP	**0.319**	$\leq$ 2.22e-16	0.1184	2.2e-16
CatPath – HP	***0.178***	$\leq$ 2.22e-16	0.0130	2.2e-16
SP – PPR	-0.0967	$\leq$ 2.22e-16	0.0028	2.2e-16
HP – PPR	-0.0835	$\leq$ 2.22e-16	0.00014	0.024
SP – HP	*0.125*	$\leq$ 2.22e-16	0.0065	2.2e-16

We find that CatPath distance is most correlated with SPL. However, most importantly, human path lengths are most correlated with CatPath distances. This means that larger CatPath distances correspond to longer human paths in both ranked-difference, measured by Kendall's τ, and in path length, measured by Pearson's r. Furthermore, PPR scores have, as expected, a negative rank correlation; this is because low PPR scores typically correspond to longer network distances and therefore higher path lengths.

Figure 3 shows scatter plots for each navigation path correlation pair. These figures are an early, albeit simple, indication that CatPaths are a good model for human navigation paths. CatPaths and shortest paths are pairwise



(a) CatPath vs PPR

(b) CatPath vs Shortest Path

(c) CatPath vs Human Path

(d) Shortest Path vs PPR

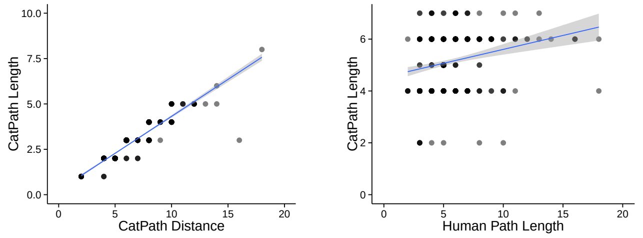
(e) Human Path vs PPR

(f) Shortest Path vs Human Path

Figure 3: Scatterplots comparing path lengths for each path navigation model. Only 500 random points are plotted in each subfigure for efficient document rendering. Slope of linear regression line indicates correlation strength. Pearson's r, Kendall's τ and corresponding p-values are shown in Table II.

compared with each other using Pearson correlation coefficient to show that these sets are not correlated. A value of 1 indicates highly correlated sets while a value of 0 implies no correlation. These results were significant with p-value ≈ 0. Additionally, we subjected pairs of CatPaths, shortest paths, and human paths to Kendall's $\tau-$test to calculate the correlation coefficient and significance values (p-values) as a relative measure of linear dependency of two variables.

Exploring these plots a bit closer we can see, *e.g.*in Figure 3b, evidence that our model of human navigation computes CatPaths with higher path length compared to the optimal node count. This figure appears sparse on the surface because many of the node hops are distributed very tightly on the shortest path range of 1 and 3. The figure shows a large and significant number of shortest paths of length 1, 2 and 3. Where as CatPaths plotted against HPL, e.g. in Figure 3c, have a wider range. A close look at the actual values highlights the evidence that the mean of the path length from CatPath approximates that of HPL. Lastly, in Figure 3f we see again that SPL, which we interchangeably equate with the optimal path length) is tightly bounded between 1 and 3, inclusive, but human paths have a wider range of values. This empirical evidence tells us that humans traverse through a higher node count space than the network's optimal shortest path. Personalized PageRank as a function of CatPath, shortest path, and HPL, e.g. in Figures 3a,3d, and 3e, collectively show that local importance or influence of each node in a chain has about the same ranking and is independent of path length type.



(a)

(b)

Figure 6: CatPath length (*i.e.*, p-path length) as a function of: (a) the CatPath distance (*i.e.*, sum of weights), and (b) HPL. Note that only 500 random points were plotted in these figures for efficient document rendering.

CatPath distance is defined as the sum of their path's edge weights. For example, in Figure 2, although the final p-path has a length of 6, the CatPath distance is 15 (not illustrated). In general, the total weight of a CatPath must be at least twice as long as its length because each c-path has a minimum length of 2. Figure 6 shows the correlation of the p-path lengths as a function of the CatPath distance (6a) ($\tau = .867, r^2 = 0.80, p < .001$) and as a function of their corresponding HPL (6b) ($\tau = 0.296, r^2 = 0.049, p < .001$). A strong correlation between weight and length is obvious,

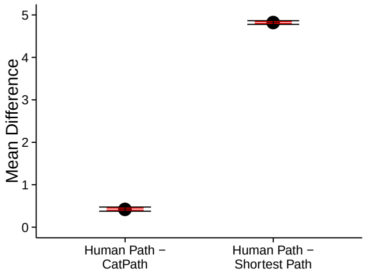
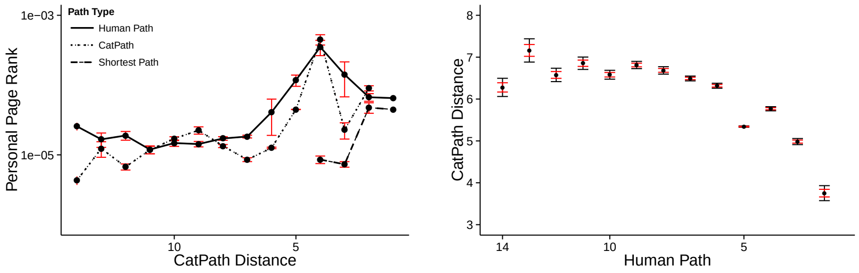


Figure 4: Mean difference between human paths and CatPath or shortest path with standard error (red) and 95% confidence intervals (black).



(a)

(b)

Figure 5: CatPath distance-to-go as a function of (a) PPR, and (b) human path length-to-go with standard error (red) and 95% confidence intervals (black)

but the correlation between human paths and p-path lengths is only slightly positive.

If we compute the difference in path lengths for all paths, we view path similarity from a different perspective. Figure 4 shows the mean difference in path lengths between human navigation paths and CatPath distances as well as human navigation paths and shortest paths. Our study of the completed paths, in the Wiki Game, yields median values for shortest and human paths centered around 2 and 5 respectively. This result is consistent with other studies of the Wiki Game (for won or completed games) [17] that find actual paths to be more than 1 hop away from the optimal path length. The PPR score does not actually generate a path, therefore path length measurements are not available for PPR.

Rather than total path length, a better way to compare paths is by observing how they change as the human navigator moves closer to the target. Figure 5a shows the mean PPR scores as a function of how CatPath and shortest path navigate through the network. We find that as humans navigate towards the target their PPR scores steadily rise to an average peak 4 steps away from the target before dropping as they finalize their path. Recall that these PPR scores are driven from the starting node, not the target node; thus, the initial rise in scores indicate that humans navigate to highly-connected (*i.e.*, high PageRank) yet topically specific nodes before finalizing their path through other nodes. The peak of human paths in Figure 5a is echoed in results by West and Leskovec [7] and indicate a tipping point between nodes that are relevant to the source-node (left-side of the peak) and nodes that are relevant to the destination-node (right-side of the peak). Most importantly, we find that the CatPath model generates paths that possess this important property.

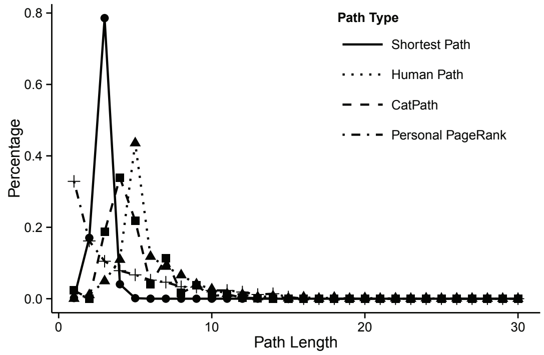


Figure 7: Distribution of path lengths for various models. PPR does not result in a path length; rather the PPR score between the source and the destination nodes discretized into 30 equal size bins for comparison-sake.

In the same spirit at the above results, Figure 5b shows the CatPath distance-to-go as a function of human-path lengths. We find that CatPath distances correlate with human path lengths up to a distance of about 7. After this point, the CatPath distances do not continue rise with human path lengths. These results indicate that CatPath distances are better at modeling short human navigation paths than long, meandering human paths. This is because the CatPath model does not account for randomness or missteps that are common in human navigation, and because longer human paths are likely caused by missteps or random-guessing, the non-stochastic CatPath model not correlate well with longer human paths.

Figure 7 show the distribution of path lengths for the various models. The HPL distributions shown here are similar in size and shape to the human path results of West and Leskovec, however the shortest path distribution described in Figure 7 is skewed leftward as compared the smaller graph used by West and Leskovec [7]. This is likely due to densification laws that govern graph sizes; that is, as graphs grow in size their average shortest path size shrinks [18]. We find that the distribution of CatPath lengths and lengths of human paths is very similar, even in the tail of the distribution. Scores for SPL and PPR (discretized into 30 bins) do not match the size and shape of human paths.

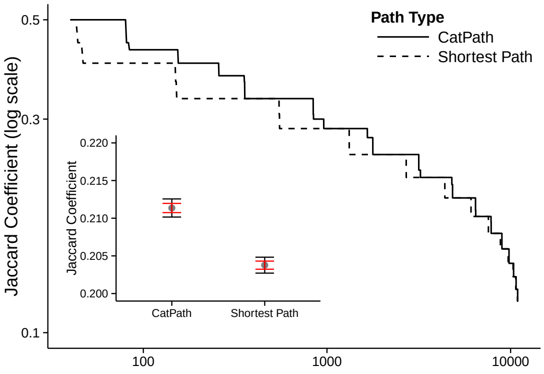


Figure 8: Distributions of Jaccard coefficients of CatPath traces and shortest paths compared to human path traces, as well as mean Jaccard coefficients with standard error (red) and 95% confidence intervals (black) (inset). Higher is better.

Path lengths, distances and differences show the size and shape of the path topology. Another way to compare the models is by looking at the actual node labels (*e.g.*, the title of the Wikipedia article) and finding the amount of overlap between various paths. For example, if a human path contained the p-nodes $u_p \rightarrow x_p \rightarrow y_p \rightarrow z_p \rightarrow v_p$, the corresponding CatPath contained $u_p \rightarrow x_p \rightarrow a_p \rightarrow z_p \rightarrow v_p$ and the shortest path contained $u_p \rightarrow b_p \rightarrow v_p$ then we would conclude that the CatPath was more similar to the human path because it has x_p and z_p in common, whereas the human path and shortest path have no nodes in common[2].

Figure 8 shows the distribution of Jaccard coefficients as well as the mean Jaccard coefficients (inset) comparing human paths to shortest paths and the CatPath model respectively. We find that the CatPath model generates higher Jaccard coefficients (*i.e.*, a larger normalized overlap) with the human path traces than do shortest paths. This indicates that the actual nodes selected by CatPath overlaps those chosen by humans more frequently than alternate models.

IV. Conclusions

In this paper we develop the CatPath model for human navigation. This model is based on the observation that human paths are typically longer than shortest-path estimates, are based on local perceptions of conceptual relatedness, and have a peculiar tipping point pattern that indicates when a human is heading towards the destination rather than away from the source.

We report the results of a suite of experiments on traces of human paths through Wikipedia via the Wiki Game and conclude that paths generated by the CatPath model are more similar to human paths than alternatives.

V. Acknowledgments

We owe a large debt to Alex Clemesha, the creator of the Wiki Game, and thank him for his willingness to share human navigation paths. This work is sponsored by the AFOSR grant AFOSR FA9550-15-1-0003, and the John Templeton Foundation grant FP053369-M.

References

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998.

[2] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *ICDM*. IEEE, 2006.

[3] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442.

[4] H. Zhu and Z.-X. Huang, "Navigation in a small world with local information," *Phys. Rev. E*, vol. 70, p. 036117, Sep 2004.

[5] J. A. Capitan, J. Borge-Holthoefer, S. Gomez, J. Martinez-Romo, L. Araujo, J. A. Cuesta, and A. Arenas, "Local-based semantic navigation on a networked representation of information," *PLoS ONE*, vol. 7, no. 8, p. e43694, Aug. 2012.

[6] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, no. 1, pp. 61–67, 1967.

[7] R. West and J. Leskovec, "Human wayfinding in information networks," in *WWW*, 2012, pp. 619–628.

[8] 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*, vol. 9, no. 7, p. e102070, 2014.

[9] P. Ganesan, H. Garcia-Molina, and J. Widom, "Exploiting hierarchical domain structure to compute similarity," *ACM Trans. Inf. Syst.*, vol. 21, no. 1, pp. 64–93, Jan. 2003.

[10] L. Adamic and E. Adar, "How to search a social network," *Social Networks*, vol. 27, no. 3, pp. 187 – 203, 2005.

[11] C. Trattner, P. Singer, D. Helic, and M. Strohmaier, "Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks," in *i-KNOW*, New York, NY, USA: ACM, 2012, pp. 14:1–14:8.

[12] D. Helic, M. Strohmaier, M. Granitzer, and R. Scherer, "Models of human navigation in information networks based on decentralized search," in *HyperText*, 2013, pp. 89–98.

[13] B. Shi and T. Weninger, "Mining interesting meta-paths from complex heterogeneous information networks," in *ICDM-MODAT*, 2014.

[14] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin, "Powergraph: Distributed graph-parallel computation on natural graphs." *OSDI*, vol. 12, no. 1, p. 2, 2012.

[15] F. Takes and W. Kosters, "Mining user-generated path traversal patterns in an information network," in *WI/IAT*, vol. 1, Nov 2013, pp. 284–289.

[16] D. Benz, A. Hotho, S. Stützer, and G. Stumme, "Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge," in *ACM WebSci*, 2010.

[17] F. W. Takes and W. A. Kosters, "Mining user-generated path traversal patterns in an information network." in *Web Intelligence*. IEEE Computer Society, 2013, pp. 284–289.

[18] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *SIGKDD*, 2005, pp. 177–187.

[2] source and destination nodes are always in common in all experiments