HypTrails: A Bayesian Approach for Comparing Hypotheses About Human Trails on the Web

Philipp Singer
GESIS & Graz University of Technology
philipp.singer@gesis.org

Denis Helic
Graz University of Technology
delic@tugraz.at

Andreas Hotho
University of Würzburg
hotho@informatik.uni-wuerzburg.de

Markus Strohmaier
GESIS & University of Koblenz-Landau
markus.strohmaier@gesis.org

ABSTRACT

When users interact with the Web today, they leave sequential digital trails on a massive scale. Examples of such human trails include Web navigation, sequences of online restaurant reviews, or online music play lists. Understanding the factors that drive the production of these trails can be useful for, e.g., improving underlying network structures, predicting user clicks or enhancing recommendations. In this work, we present a general approach called HypTrails for comparing a set of hypotheses about human trails on the Web, where hypotheses represent beliefs about transitions between states. Our approach utilizes Markov chain models with Bayesian inference. The main idea is to incorporate hypotheses as informative Dirichlet priors and to leverage the sensitivity of Bayes factors on the prior for comparing hypotheses with each other. For eliciting Dirichlet priors from hypotheses, we present an adaption of the so-called (trial) roulette method. We demonstrate the general mechanics and applicability of HypTrails by performing experiments with (i) synthetic trails for which we control the mechanisms that have produced them and (ii) empirical trails stemming from different domains including website navigation, business reviews and online music played. Our work expands the repertoire of methods available for studying human trails on the Web.

Categories and Subject Descriptors: H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—Web-based interaction

Keywords: Human Trails; Paths; Sequences; Hypotheses; Web; Sequential Human Behavior; Markov Chain; Bayesian Statistics

1. INTRODUCTION

The idea of human trails in information systems can be traced back to early work by Vannevar Bush (“As We May Think” [9]), in which he described a hypothetical system called Memex. Bush hypothesized that human memory operates by association, with thoughts defined by internal connections between concepts. The Memex itself was intended as users’ extension of their memory, where common associative trails between documents can be stored. Accessed and shared. Eventually, Bush’s ideas led to the concept of Hypertext [27] and the development of the World Wide Web [4].

Today, the Web facilitates the production of human trails on a massive scale; examples include consecutive clicks on hyperlinks when users navigate the Web, successive songs played in online music services or sequences of restaurant reviews when sharing experiences on the Web. Understanding such human trails and how they are produced has been an open and complex challenge for our community for years. A large body of previous work has tackled this challenge from various perspectives, including (i) modeling [13, 31, 35], (ii) regularities and patterns [20, 42, 43] and (iii) cognitive strategies, finding that, for example, humans prefer to consecutively choose semantically related nodes [36, 45], humans participate in partisan sharing [2] or users benefit from following search trails [47]. In this paper, we are interested in tackling an important sub-problem of this larger challenge.

Problem. In particular, we take a look at the problem of expressing and comparing different hypotheses about human trails given empirical observations. We define a trail as a sequence of at least two successive states, and hypotheses as beliefs about transitions between states. An intuitive way of expressing such hypotheses is in the form of Markov transitions and our beliefs in them. For example, we might have various hypotheses about how humans consecutively review restaurants on Yelp. Figure 1 (a-c) shows three exemplary hypotheses for transitions between five restaurants (A-E) in Italy, and actual empirical transitions (d). The uniform hypothesis in Fig 1(a) expresses a belief that all transitions are equally likely (a complete digraph). In Fig 1(b), the geo hypothesis expresses a belief that humans prefer to consecutively review restaurants on Yelp. In Fig 1(c), the self-loop hypothesis expresses a belief that humans repeatedly review the same restaurant without ever reviewing another one. Other hypotheses are easily conceivable. What is difficult today is expressing and comparing such hypotheses within a coherent research approach. Such an approach would allow to make relative statements about the plausibility of different hypotheses given empirical data about human trails.

Objectives. We thus tackle the problem of comparing a set of hypotheses about human trails given data. We present a Bayesian approach—which we call HypTrails1—that provides a general solution to this problem.

1 Portmanteau for Hypertext/theses Trails
2. THE HYPTRAILS APPROACH

We aim to produce a partial ordering \( O \) over a set of hypotheses \( H = \{H_1, H_2, \ldots, H_n\} \). We base the partial order on the plausibility of hypotheses given data \( D \). Each hypothesis \( H \) describes beliefs about common transitions between nodes while data \( D \) captures empirically observed human trails. A hypothesis \( H \) can be expressed by an adjacency matrix \( Q \) where transitions \( q_{i,j} \) with strong belief receive larger values than those with lower belief.

For generating the partial ordering \( O \), our HypTrails approach resorts to Bayesian inference utilizing a Markov chain model. We incorporate a hypothesis \( H \) as informative Dirichlet priors into the inference process. For eliciting Dirichlet priors \( \text{Dir}(\alpha) \) from a given hypothesis \( H \) expressed as matrix \( Q \)—i.e., for setting corresponding hyperparameters \( \alpha_{i,j} \)—HypTrails uses an adaption of the so-called (trial) roulette method. The partial ordering \( O \) is achieved by calculating marginal likelihoods \( P(D|H) \) (weighted averages of likelihood, where the weights come from the parameters’ prior probabilities) for competing hypotheses \( H \) which we then can compare with each other by determining Bayes factors \( B \).

2.2 Bayesian Markov Chain Modeling

HypTrails is based on Bayesian Markov chain modeling. In the following, we only cover those fundamentals that are directly related to our approach, and point the reader to previous work [35, 38] for a more detailed treatise of the topic.

Markov chain definition. A Markov chain model represents a stochastic system that models transitions between states from a given state space \( S = \{s_1, s_2, \ldots, s_m\} \) with \( m = |S| \) (e.g., the distinct restaurants of our example in Figure 1). It amounts to a sequence of random variables \( X_1, X_2, \ldots, X_n \). This random process is usually memoryless (the so-called Markov property, first-order) meaning that the next state only depends on the current state and not on a sequence of preceding states. Note though that Markov chain models can also be extended to incorporate higher orders; see Section 5 for a discussion. We can define the Markov property as:

\[
P(X_{t+1} = s_j|X_1 = s_{i_1}, \ldots, X_{t-1} = s_{i_{t-1}}, X_t = s_i) = P(X_{t+1} = s_j|X_t = s_i) = p_{i,j}.
\]

Markov chain models have been established as a robust method for modeling human trails on the Web in the past (e.g., [14, 35, 43]),

2.1 Problem Definition & Approach

We start with defining the problem setting and giving a short overview of the proposed approach in Section 2.1. We proceed with explaining the fundamentals of our approach based on Bayesian Markov chain modeling in Section 2.2 where we also emphasize our main idea of incorporating hypotheses as Dirichlet priors and leveraging the sensitivity of Bayes factors for comparing hypotheses with each other. In Section 2.3 we thoroughly discuss the process of eliciting Dirichlet priors from scientific hypotheses.

Approach & Methods. The HypTrails approach utilizes a Markov chain model for modeling human trails and Bayesian inference for comparing hypotheses. The main idea is to (i) let researchers express hypotheses about human trails as adjacency matrices which are then used for (ii) eliciting informative Dirichlet priors using an adapted version of the (trial) roulette method. Finally, the approach (iii) leverages the sensitivity of Bayes factors for comparing hypotheses with each other. We experimentally illustrate our approach by studying synthetic datasets with known mechanisms which we then express as hypotheses. We demonstrate the general applicability of HypTrails by comparing hypotheses for empirical datasets from three distinct domains (Wikigame, Yelp, Last.fm).

Contributions. Our main contribution is the presentation of HypTrails, a general approach for expressing and comparing hypotheses about human trails. While the basic building blocks of HypTrails are well established (Markov chains, Bayesian inference), we combine, adapt and extend them in an innovative way that facilitates intuitive expression and elegant comparison of hypotheses. In particular, our adaption of the (trial) roulette method represents a simple way of eliciting priors for Markov chain modeling. We demonstrate the applicability of our framework in a series of experiments with synthetic and real-world data. Finally, to facilitate reproducibility and future experimentation, we make an implementation of HypTrails openly available to the community.

Structure. We present our approach in Section 2. Section 3 describes the synthetic and empirical data analyzed; corresponding experiments are presented in Section 4. We discuss our work in Section 5, present related work in Section 6 and conclude in Section 7.
specifically focusing on human navigational trails (e.g., [6, 31]) with Google’s PageRank being the most prominent example [7]. Hence, the Markov chain model is a natural and intuitive choice for our approach as it lets us explicitly model human trails with a dependence of the next state on the current state. We also consider hypotheses as beliefs about transitions without memory.

A Markov model is usually represented by a stochastic transition matrix $P$ with elements $p_{i,j} = P(s_j|s_i)$ which describe the probability of transitioning from state $s_i$ to state $s_j$; the probabilities of each row sum to 1. The elements of this matrix are the parameters $\theta$ that we want to determine. For doing so we resort to Bayesian inference.

**Bayesian inference.** Bayesian inference refers to the Bayesian process of inferring the unknown parameters $\theta$ from data; it treats data and model parameters as random variables. For a more detailed discussion of Bayesian inference please refer to [35, 38]. Following Bayes’ rule, the posterior distribution of parameters $\theta$ given data $D$ and hypothesis $H$ is then defined as:

$$
P(\theta|D, H) = \frac{P(D|\theta, H)P(\theta|H)}{P(D|H)}
$$

The likelihood function describes the likelihood that we observe data $D$ with given parameters $\theta$ and hypothesis (model) $H$. The prior reflects our belief about the parameters before we see the data or, more technically, the prior encodes our hypothesis $H$. Thus, we use the prior as a representation for different hypotheses about human trails. More precisely, we model the data with different, mostly informative, priors. We use the conjugate prior of the categorical distribution as the prior of each row of the transition matrix $P$; i.e., the Dirichlet distribution $Dir(\alpha)$. The hyperparameters $\alpha$ represent our prior belief of the parameters and can be seen as a vector of pseudo counts $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_m]$. Given such a prior, the posterior distribution represents a combination of our prior belief and the actual data that we observe. For each row $i$ of $P$ we now have a posterior in the form of $Dir(n_{i1} + \alpha_1, \ldots, n_{im} + \alpha_m)$ where $n_{ij}$ are the actual transition counts of the data between states $s_i$ and $s_j$ and $\alpha_{i,j}$ are the prior pseudo counts assigned to this transition. We provide a thorough description of how we elicit the needed Dirichlet priors from expressed hypotheses by researchers in Section 2.3.

**Comparing hypotheses.** Finally, the marginal likelihood (which we can also call evidence) expresses the probability of the data given a hypothesis $H$ and plays a crucial role for comparing hypotheses; it is defined as follows (for derivation please consult [35, 38]):

$$
P(D|H) = \prod_i \frac{\Gamma(\sum_j \alpha_{i,j})}{\prod_j \Gamma(\alpha_{i,j})} \prod_j \Gamma(n_{ij} + \alpha_{i,j}) / \Gamma(\sum_j (n_{ij} + \alpha_{i,j}))
$$

Note that the hyperparameters $\alpha_{i,j}$ differ for various hypotheses $H$ as we express them via different Dirichlet priors; the actual transition counts $n_{ij}$ are the same for each hypothesis. For comparing the plausibility of two hypotheses, we resort to Bayes factors [21, 44]. Bayes factors are representing a Bayesian method for model comparison that include a natural Occam’s razor guarding against overfitting. In our case, a model represents a hypothesis at interest with each having different priors with different hyperparameters that express corresponding beliefs. For illustrative purposes, we are now interested in comparing hypotheses $H_1$ and $H_2$ where $H_1, H_2 \in H$, given observed data $D$. We can define the Bayes factor—note that we apply unbiased comparison assuming that all hypotheses are equally likely a priori—as follows:

$$
B_{1,2} = \frac{P(D|H_1)}{P(D|H_2)}
$$

$P(D|H)$ is the marginal likelihood (evidence) defined in Equation 3 and the Bayes factor can be seen as a summary of the evidence provided by the data in favor of one scientific hypothesis over the other. HypTrails is not only suited for comparing two hypotheses with each other, but rather a set of hypotheses $H = \{H_1, H_2, \ldots, H_n\}$. For determining the partial order $O$ over $H$, we order the evidences that data $D$ provides in favor of hypotheses $H$; i.e., by ordering $P(D|H)$ using a less-than-equal binary relation. However, ordering the evidences is not enough as we need to check the significance of their ratios which we tackle by calculating Bayes factors. In case that the significance is not present, we consider two hypotheses as being equal. For determining the strength of the Bayes factor we resort to Kass and Raftery’s [21] interpretation table.

**Leveraging the sensitivity of Bayes factors.** Throughout this section we have described our main idea of incorporating hypotheses in the form of informative Dirichlet priors into the inference process. We leverage marginal likelihoods and Bayes factors for making informed decisions about the relative plausibility of given hypotheses. Usually, one common critique of Bayes factors is that they are highly sensitive with regard to the choice of the prior [21]. In contrast, posterior measures ignore the influence of the prior the more data one observes which is why they are more ignorant to the choice of prior [39]. In our approach, we exploit the sensitivity of Bayes factors on the prior as an elegant solution to the problem of comparing hypotheses. As we express our different hypotheses in the form of priors, we are explicitly interested in using a measure that is sensitive to the choice of priors and hence, can give us insights into the relative plausibility of each hypothesis. Thus, in this case, marginal likelihoods and Bayes factors are an appropriate measure for comparing scientific hypotheses [39].

### 2.3 Eliciting Dirichlet Priors

This section explains in greater detail how we can express the hypotheses about human trails and how HypTrails elicits proper informative Dirichlet priors from these. First, we illustrate how the prior influences the evidence by studying several toy examples. Next, we present an intuitive way of eliciting the Dirichlet priors by introducing an adaption of the so-called (trial) roulette method.

**Understanding influence of priors.** In Section 2.2, we discussed that we use the sensitivity of Bayes factors with regard to the choice of prior (i.e., determined via marginal likelihoods) as a feature (or solution) rather than a limitation. It allows us to model hypotheses in the form of prior distributions which then can be compared by corresponding Bayes factors. But how exactly does the prior influence evidence (marginal likelihood)? Note that the posterior probability (see Equation 2) is a combination of our prior belief (pseudo counts) and the data we observe (transition counts) which is why both influence the evidence (see Equation 3).

To illustrate the influence, we apply several toy priors to human trail data. The choice of data is secondary and we observe the same behavior regardless of the underlying data; in this case we exemplary use navigation data (Wikigame dataset introduced in Section 3). First, we apply a *uniform prior*; i.e., $\alpha$ has the same value for each row $i$ and element $j$: $\alpha_{i,j} = 1 + c, \forall i, j$. By ranging the constant $c$ over $0, 1, 3, 5, 10, 20$, we observe decreasing evidence (Figure 2(a)) which is not surprising as the uniform pseudo counts do not mirror the observed transition counts well. Technically, with increased $c$
the Dirichlet prior concentrates more and more of its probability mass on a uniform distribution of parameters, and thus, the weights of alternative parameter configurations become smaller. However, the likelihood is larger for the alternative parameter configurations (coming from the data) and this results in a smaller weighted average aligned prior.

On the other hand, if we provide some form of an empirically aligned prior as \((n_{ij} > 0 \rightarrow \alpha_{ij} = 1 + c) \land (n_{ij} = 0 \rightarrow \alpha_{ij} = 1), \forall i, j\) we end up with a larger evidence the larger \(c\) is as we can see in Figure 2(b). This is because we actually increase the pseudo counts of transitions that we also observe in our data while we keep the pseudo counts for non-observed transitions at 1. In this case, we concentrate the prior probability mass on the parameter configuration that is very well aligned with the observations. As a consequence we give more weight for parameter configurations where the likelihood is anyhow large and this increases the evidence.

Finally, we illustrate the behavior of a toy prior that expressed an empirically opposing prior in the form of \((n_{ij} = 0 \rightarrow \alpha_{ij} = 1 + c) \land (n_{ij} > 0 \rightarrow \alpha_{ij} = 1), \forall i, j\). In this example we intentionally set the prior pseudo counts to the opposite of what the actual data tells us. We assign low prior pseudo counts \((1)\) to elements with large observed transition counts while we incrementally increase the pseudo counts of transitions that we do not observe in our data. As expected, Figure 2(c) shows that the evidence decays as we increase \(c\). Technically, we assign the greatest weights for parameter configurations with the smallest likelihoods resulting in a steep decay of evidence with increasing values of \(c\).

Note that \(c = 0\) results in the same evidence for all three toy priors as in all cases \(\alpha_{ij} = 1, \forall i, j\). This refers to a Dirichlet prior that is uniformly distributed over the simplex which is a special form of a uniform prior and which we will refer to as a flat prior.

These toy examples show that if the prior is well aligned with data, the evidence is rising with the strength of the prior. The evidence is the largest if the prior and the likelihood are concentrated over the same parameter regions and it is the lowest if they concentrate on different regions [48]. Hence, we want to choose an informative prior that captures the same regions as the likelihood. This leads to the observation that if the prior choice represents a valid hypothesis about behavior producing human trails on the Web, the evidence should be larger than a uniform prior, or an unlikely hypothesis prior with an equal amount of pseudo counts. Thus, for fair comparison, we want to compare hypotheses with each other that exhibit the same amount of pseudo counts assigned. Next, we elaborate the (trial) roulette method for eliciting Dirichlet priors from hypotheses.

Figure 2: Understanding influence of priors. This figure shows how the choice of prior pseudo counts influences evidence; we apply several toy priors to navigation data (Wikigame, see Section 3). In (a) we use a uniform Dirichlet prior which means that for each row \(i\), \(\alpha\) has the same value for each element: \(\alpha_{ij} = 1 + c, \forall i, j\). By increasing the constant \(c\) (x-axis), we can observe that the evidence (y-axis) is decreasing; the largest evidence is at \(c = 0\). By using an empirically aligned prior (b) as \((n_{ij} > 0 \rightarrow \alpha_{ij} = 1 + c) \land (n_{ij} = 0 \rightarrow \alpha_{ij} = 1), \forall i, j\), we end up with a larger evidence the larger \(c\) is. Finally, in (c) we intentionally set "bad" prior counts for the \(\alpha\) values via \((n_{ij} = 0 \rightarrow \alpha_{ij} = 1 + c) \land (n_{ij} > 0 \rightarrow \alpha_{ij} = 1), \forall i, j\) showing that the evidence becomes smaller as we increase \(c\). The results indicate that the more a hypothesis is aligned with empirical data, the larger the evidence is, and vice versa.

(Trial) roulette method. Our approach requires to define the hyperparameters of prior Dirichlet distributions by setting the pseudo counts \(\alpha_{ij}\) given the hypothesis at interest. However, the process of eliciting prior knowledge is no trivial problem and requires careful steps (see [18, 28] for a discussion). As a solution, we present an adaption of the so-called (trial) roulette method which was originally proposed in [19] and further discussed in [16, 28]. It is a graphical method that allows experts to express their subjective belief by distributing a fixed set of chips (think about casino chips you set on a roulette table) to a given grid (e.g., bins representing result intervals). The number of chips assigned to an element of the grid then reflect the experts’ belief in the specific bin.

In our work we adapt the (trial) roulette method. The grid can be understood as a matrix \(Q\) where each element \(q_{ij}\) of the grid represents the belief of a given hypothesis about the transition from state \(s_i\) to state \(s_j\). Values \(q_{ij}\) are set by researchers for expressing a hypothesis. They need to be positive values and larger values indicate stronger belief in a given transition. The prior of each row of the transition matrix \(P\) of the Markov chain model is defined as a Dirichlet distribution (cf. Section 2.2) with hyperparameters (pseudo counts) \([\alpha_1, \alpha_2, \ldots, \alpha_q]\) which we want to set given the hypothesis. Concretely, we want to automatically distribute a number of chips to the given pseudo counts (another grid) according to the values provided in matrix \(Q\) expressing a hypothesis \(H\). We define the overall number of chips to distribute for a given hypothesis as:

\[
\beta = \frac{m}{m^2} + \frac{km^2}{m^2 + km^2}
\]  

where \(m = |S|\) and \(m^2\) amounts to the flat prior—i.e., we assign the same number of pseudo counts \((1)\) to each transition—which is why the number of uniformly assigned chips is equal to the overall number of parameters of the Markov chain model. Additionally, we distribute \(k \cdot m^2\) informative pseudo clicks for the given hypothesis, where \(k\) describes a weighting factor for the informative part. The larger \(k\), the more we concentrate the Dirichlet distributions according to a hypothesis at interest—see Section 5 for a discussion.

By and large, the goal of our adaption of the (trial) roulette method is to not only give researchers an intuitive way of expressing their hypotheses as matrices \(Q\), but also to elicit informative Dirichlet distributions according to the values \(q_{ij}\) of \(Q\). Next, we want to illustrate the process of expressing a hypothesis and assigning prior pseudo counts via the example shown in Figure 3 using the (trial) roulette method. Let us again focus on human trails over reviewed restaurants in Italy (see Figure 1) and illustrate the method using the
We begin with a matrix expressing a researcher’s hypothesis about human trails in (a)—in this case the exemplary geographic hypothesis given in Figure 1(b). For this visualization, we assume that we set $k = 1$ leading to $\beta = m^2 + m^2$ chips we want to distribute. The following steps are necessary:

(a) Expressing the hypothesis. Researchers start with expressing the hypothesis matrix $Q$ with elements $q_{ij}$ that capture the belief about transitions of underlying human trails; the matrix can be seen in Figure 3(a). In this example, we have five states (restaurants)—i.e., $S = A, B, C, D, E$ and $m = 5$ leading to $\beta = 50$ chips to distribute—and we express our geographic hypothesis about common transitions with values between 0 and 1. The precondition is that only positive values are used and larger values ($q_{ij}$) always express stronger beliefs compared to smaller values. In this case, the closer a value is to 1, the closer two restaurants are geographically, and the more we believe in corresponding transitions. As can be seen in Figure 1(b), both restaurant pairs B-C and E-D are the closest in geographical terms which is why we lay the strongest belief in these symmetrical transitions; concretely, we set $q_{BC} = 1.0, q_{CB} = 1.0, q_{DE} = 1.0$ and $q_{ED} = 1.0$. The next closest restaurant pair is A-C which is why we also have strong beliefs in humans consecutively reviewing restaurant A after C and vice versa; we set $q_{AC} = 0.9$ and $q_{CA} = 0.9$. Finally, we set $q_{AB} = 0.7$ and $q_{BA} = 0.7$ as we also have some (lower) belief that humans consecutively review given restaurants. We set all other transitions to zero as we believe that given restaurants are too far away. One matrix $Q$ represents one general hypothesis $H$ about human trails—in a more realistic scenario one would want to express several of such matrices (such as for the other hypotheses given in Figure 1). Also note that we hand-pick values for this example and in a more rigorous investigation one would potentially use an automatic method for determining them (as we do in Section 4). The next steps are automatically performed by HypTrails for eliciting Dirichlet priors from such matrices, more specifically by the adapted (trial) roulette method.

(b) Initial uniform distribution. The (trial) roulette method starts with assigning uniform chips to each transition which can be seen as obtaining a flat prior ($\alpha_{ij} = 1, \forall i, j$) and accounts to the left part ($m^2$) of Equation 5. The updated prior (i.e., hyperparameters for the Dirichlet distributions) can be seen in Figure 3(b) with black bars; all elements where one chip is assigned to are marked bold and black in Figure 3(a). By subtracting the distributed number of chips from $\beta$ we have $\beta = 50 - 25 = 25$ chips left for the informative part described next.

(c) Informative distribution. Matrix $Q$ gets normalized and then multiplied by the number of chips left: $Q = \frac{Q}{||Q||} \cdot \beta$ where $||Q||_1$ is the $\ell_1$-norm and $\beta = 25$. The resulting matrix can be seen in Figure 3(c1). The method assigns as many chips to elements of the prior as the integer floored values of $Q$ specify. So e.g., $q_{AB} = 2.43$ and $q_{AB} = 2$ leading to $q_{AB} = 2$ whereas $q_{BD} = 0$ which is why the pseudo count for this transition is not increased. Overall, the method distributes 22 more chips marked bold and blue in Figure 3(c1) leading to $\beta = 25 - 22 = 3$ chips left; the updated prior distributions (new chips marked blue) can be seen in Figure 3(c2).

(d) Remaining informative distribution. Finally, the method subtracts the integer floored values from $Q$ leading to the matrix illustrated in Figure 3(d1) calculated by $Q = Q - \lfloor Q \rfloor$. It now needs to distribute the chips left (three in this case) according to the remaining values in $Q$. The method accomplishes this by ranking the values in descending order and assigning one chip to each element until none is left, starting from the largest and ending at the smallest. In case of a tie the ranking for the ties is produced in random fashion; hence, in this case $q_{AE}$ does not receive one more chip. We mark the elements that receive one further chip bold and red in Figure 3(d1) and update our prior pseudo counts as can be seen in Figure 3(d2) also in red color. Now, the (trial) roulette method has no chips left and is finished.
The final chip assignment shown in Figure 3(d2) represents the prior (hypothesis). In detail, each row corresponds to a Dirichlet distribution with corresponding pseudo counts (hyperparameters) αc,j (e.g., αc,B = 5). By proceeding, our HypTrails approach now uses these Dirichlet priors for Bayesian Markov chain modeling inference as described in Section 2.2. Concretely, in combination with the transitions qij, observed from data they influence the marginal likelihood calculated as defined in Equation 3. See Figure 2 for a visualization of how the prior influences the evidence. By repeating the trial roulette method and evidence calculation for each hypothesis at interest, we can determine corresponding evidences and Bayes factors for comparing hypotheses with each other.

3. DESCRIPTION OF DATASETS

In this section, we introduce both synthetic as well as empirical datasets which we consider for our experiments. We produce synthetic data consisting of simulated human trails—in this case navigational trails—with known mechanisms from a generated (i.e., artificial) network. The introduced empirical data stems from three real-world datasets from different domains. The state space S investigated is always defined by the distinct elements the trails traverse over; e.g., if we observe trails over five distinct restaurants being reviewed (see Figure 1) we consider these five for the state space. As described in [35], we additionally add a reset state to the trails and state space of each dataset to ensure ergodicity.

3.1 Synthetic Datasets

We start by generating a directed random network using a generalized version of Price’s preferential attachment scale-free network model [3, 32]. The network generation algorithm starts with a clique containing 11 nodes and proceeds to add nodes with an out-degree of 10 leading to an overall network size of 10,000 nodes. These parameters are arbitrary and could be set differently. Next, we simulate three different kinds of navigational trails, each consisting of exemplary 1,000 trails of length 5, as follows:

- **Structural random walk.** For each trail we start at a random node of the network and perform a random walk through the network. The walker chooses the next node by randomly selecting one out-going link of the current node.

- **Popularity random walk.** Again, the walker starts at a random node of the network, but now selects the next node by choosing the out-link according to the target’s in-degree. The probability of choosing the next node according to its in-degree is defined as max (deg (i)/10), which is then used for selecting the next node so as to make the process less stochastic. This is aimed at averting too long loops that would happen with simple greedy selection.

- **Random teleportation.** Again, we start with a random node in the network for each trail. However, we now completely ignore the underlying topological link network and simply randomly choose any other node of the network—i.e., teleporting through the network.

3.2 Empirical Datasets

For our experiments we also consider three different empirical datasets which are described next.

- **Wikigame dataset.** First, we study navigational trails over Wikipedia pages that are consecutively visited by humans. The dataset is based on the online game called Wikigame (thewikigame.com) where players aim to navigate to a given Wikipedia target page starting from a given Wikipedia start page using Wikipedia’s link structure only. All start-target pairs are guaranteed to be connected in Wikipedia’s topological link network and users are only allowed to click hyperlinks and use the browser’s buttons such as refresh, but not use other features such as the search field. In this article, we study trails collected from users playing the game between 2009-02-17 and 2011-09-12. Overall, the dataset consists of 1,799,015 trails—where each trail represents the consecutive websites visited by one user for one game played—through Wikipedia’s main namespace including 360,417 distinct pages with an average trail length of around 6. We use corresponding textual and structural Wikipedia article data for hypotheses generation. In particular, we use the Wikipedia dump dated on 2011-10-07.

- **Yelp dataset.** Second, we study human trails over successive businesses reviewed by users on the reviewing platform Yelp [yelp.com]—we have used this setting as an example throughout this article (e.g., see Figure 1). For generating these trails we use a publicly available dataset [11] for generating the trails at hand focusing on data from one day (2009-01-01). Overall, the dataset consists of 275 trails—where each trail describes the subsequent review history of one single user—over 41,707 distinct businesses with an average trail length of 8. The data also includes further information about the businesses like the geographic location or category markers assigned, which we will use for hypotheses generation.

- **Last.fm dataset.** Third, we study human trails that capture consecutive songs listened to by users on the music streaming and recommendation website Last.fm [last.fm]. We use a publicly available dataset [11] for generating the trails at hand focusing on data from one day (2009-01-01). Overall, the dataset consists of 125,365 trails—where each trail describes the subsequent reviews listened to by one user on a given day—over 11,166 distinct tracks with an average trail length of 52.8. For generating hypotheses, we consult the MusicBrainz (musicbrainz.org) API as described later.

4. EXPERIMENTS

To demonstrate HypTrails and its general applicability, we perform experiments with both synthetic as well as empirical datasets (as introduced in Section 3).

4.1 Experiments with Synthetic Data

Our first experiments focus on applying HypTrails to three synthetic trail datasets, generated by the following mechanisms: teleportation, a random walk and a popularity random walk (see Section 3). We look at these three datasets and compare three corresponding hypotheses (uniform, structural, popularity) that capture the generative mechanisms of each dataset. As we know from theory, HypTrails ranks the hypothesis that best captures the underlying mechanisms as the most plausible one. Next, we introduce the hypotheses in greater detail, before we discuss the experimental results.

- **Hypotheses.** We now describe how we express the three hypotheses as matrices Q. Note that we only have to specify the hypothesis matrix Q (see Section 2.3) while the concrete pseudo count distribution for generating proper priors is handled by our approach.

- **Uniform hypothesis.** This hypothesis has the intuition that trails have been purely generated by random teleportation and all transitions are equally likely. Thus, we equally believe in each transition and set each element of Q to an equal value (here 1).

- **Structural hypothesis.** This hypothesis captures our belief that trails have been generated by (only) following the underlying topological link structure. Hence, we believe that agents would always choose a random link leading from one node to another while traversing the network. We express this by setting qij of Q to 1 if a directed link between state xi and state xj exists in the topological network.

---

4 This Wikipedia dump closely resembles the information available to players of the game for our given time period.

5 yelp.com/dataset_challenge
**Random teleportation**. This hypothesis also believes that the trails have been generated by following the links of the underlying link structure, but we have stronger beliefs in choosing large indegree nodes compared to low in-degree nodes. Hence, we set $q_{i,j}$ to $\text{deg}^+(s_j)$ if a directed link between state $s_i$ and state $s_j$ exists in the topological network.

**Results.** Figure 4 depicts the results for each hypothesis and dataset at interest. The x-axis denotes the weighting factor $k$ for the number of pseudo counts assigned (cf. Equation 5). The y-axis denotes the corresponding Bayesian evidence (marginal likelihood). For $k = 0$ the evidence is the same for all hypotheses as in that case the flat prior is uniformly distributed over the simplex and no additional informative aspect is considered. The larger $k$ gets, the more pseudo counts are assigned to the prior according to the given hypothesis and hence, the stronger our belief in specific transitions of a given hypothesis. We can compare hypotheses with each other by comparing the y-values (evidence) for the same x-values (weighting factor $k$). According to Kass and Raftery’s interpretation table of log-Bayes factors [21], we find that all differences are decisive which is why we refrain from presenting explicit Bayes factors. Hence, the larger the evidence for a given hypothesis is, the more plausible it is in comparison to the other hypotheses at interest. Across all three synthetic datasets, we can observe what we know from theory: the hypothesis that captures the underlying known mechanisms of the synthetic trails best is found to be the most plausible one. In the following we discuss the results of each dataset in detail:

**Structural random walk.** In Figure 4(a) we can see that the structural hypothesis is ranked as the most plausible one as it exhibits the highest evidences for $k > 0$. This result is as expected from theory as the trails are also produced according to a structural random walk only considering the underlying topological link network as expressed by the structural hypothesis. The reason why the popularity hypothesis is more plausible than the uniform hypothesis is because the former also incorporates structural information while the latter does not.

**Popularity random walk.** We show the results for our popularity random walk generated trails in Figure 4(b). In this case, the popularity hypothesis incorporating the in-degree (popularity) of potential structural target nodes is identified as the most plausible one as it captures the mechanisms according to which the trails have been generated.

**Random teleportation.** Finally, in Figure 4(c) we demonstrate the results for our trails generated via random teleportation. As expected, the uniform hypothesis is the most plausible one which accounts to our prior belief that all target nodes are equally likely to come next given a current node. Contrary, the structural and popularity hypotheses which both incorporate structural knowledge are less plausible hypotheses.

### 4.2 Experiments with Empirical Data

Our second kind of experiments focus on demonstrating the general applicability of the HypTrails approach by applying it to three real-world, empirical human trail datasets (Wigigame, Yelp and Last.fm) as introduced in Section 3. We compare universal as well as domain-specific hypotheses for each dataset which we describe next, before we discuss the experimental results.

**Hypotheses.** We now describe the universal and domain-specific hypotheses studied and how we express them. These are just exemplary hypotheses for illustrative purposes, researchers are completely free to formulate other / their own hypotheses accordingly.

**Uniform hypothesis.** We use the universal uniform hypothesis in a similar fashion as for our experiments with synthetic data in order to express our prior belief that each state is equally likely given a current state. Hence, we assign 1 to each element of the hypothesis matrix $Q$. We can see this hypothesis as a baseline for other hypotheses; if they are not more plausible than the uniform hypothesis, we cannot expect them to be good explanations about the behavior that is producing the underlying human trails.

**Self-loop hypothesis.** With the universal self-loop hypothesis we express our prior belief that humans never switch to another element in a trail. For example, for a navigational scenario this would mean that if a user currently is on a specific Wikipedia page, she would always just refresh the current one and never switch to another one. We set the diagonal to 1 in the corresponding hypothesis matrix $Q$ and leave all other elements zero.

**Similarity hypothesis.** This hypothesis expresses our belief that humans consecutively target nodes in trails that are in some way (e.g., semantically) related to each other. We now aim at modeling this hypothesis for all three datasets. However, due to their given nature, the similarity hypothesis differs for each dataset which is why we describe the domain-specific similarity hypotheses next:

**Wigigame similarity hypothesis.** This hypothesis believes that humans prefer to consecutively access websites that are semantically related which has been observed and hypothesized in a series of previous works (e.g., [36, 45, 46]). Using the set of Wikipedia pages that users navigate over, we use the textual information of each site
Figure 5: *Experiments with empirical data.* This figure depicts the results obtained from applying HypTrails to three different empirical human trail datasets (Wikigame (a), Yelp (b) and Last.fm (c)) for comparing a set of hypotheses. The x-axis depicts the strength (weighting factor $k$) one assigns to a given hypothesis as defined in Equation 5 ($k=0$ refers to a flat prior) while the y-axis shows the corresponding evidence (marginal likelihood) value. For simplicity, we can compare hypotheses with each other by comparing the evidence values (larger values mean higher plausibility) for the same values of $k$ as all Bayes factors are decisive.

We set elements of the hypothesis matrix $Q$ between two tracks to 1 only if they are from the same artist—the diagonal is set to zero.

**Wikigame structural hypothesis.** For Wikigame, we evaluate an additional domain-specific hypothesis that captures our prior belief that users navigate the Web (or in this case Wikipedia) primarily by using the underlying topological link structure. The corresponding hypothesis matrix $Q$ can hence be built by looking whether links between sites of our states space $S$ exist in the underlying topological link network $G$ with directed edges $E(G)$ (derived from the Wikipedia dump as stated in Section 3.2). To be precise, the values of the elements $q_{ij}$ of $Q$ are determined by the number of hyperlinks linking from page $s_i$ to page $s_j$; mostly, only one hyperlink links from one page to the other. Additionally, we set the diagonal of the matrix to 1 as users might also subsequently navigate the same page by e.g., clicking the refresh button of the browser.

**Yelp geographic hypothesis.** On Yelp, we also consider the domain-specific hypothesis that the next business a user reviews is one that is geographically close to the current one—we have used this as an example throughout this article (e.g., see Figure 1(b) or Figure 3). For doing so, we start by calculating the *haversine distance* [37] between the longitude and latitude values of all pairs of businesses. As the resulting value (in km) is smaller for geographic close businesses than for far businesses we normalize the values by dividing them by the maximum distance before subtracting them from 1. This leads to final values that range from 0 to 1 where 1 means geographically identical. We set the values of $Q$ according to the calculated values while leaving the diagonal zero.

**Last.fm date hypothesis.** Finally, we specify a hypothesis that believes that successive tracks listened to on Last.fm are close regarding their original publication date (e.g., someone prefers to only listen to 80s songs). We determine the date of a track by using the Musicbrainz API and looking for the earliest release date available. Next, we calculate the difference between dates of two songs in years—we only consider track pairs for which we can retrieve a date for both tracks through the API. Similar to the Yelp dataset, we then divide each date difference value by the maximum and subtract it from 1 giving us scores between 0 and 1 where the latter means that two tracks are originally published in the same year. We set the transition values of $Q$ according to the calculated values and leave the diagonal zero.
Results. The results for all datasets are shown in Figure 5. Again, all Bayes factors are decisive and we can simply interpret hypotheses having larger y-values (evidence, marginal likelihood) as more plausible. Across all datasets, we can identify some domain-specific hypotheses that are more plausible compared to the universal uniform hypothesis which we can see as a baseline. Hence, these hypotheses seem to capture some mechanisms well that human behavior exhibits while producing the human trails studied. Additionally, we find that throughout all datasets the universal self-loop hypothesis is the least plausible one with a small exception for the Wikigame dataset. We discuss the results for each dataset next.

Wikigame. First and foremost, the HypTrails results in Figure 5(a) show largest evidence for the domain-specific structural hypothesis. This indicates that users playing the Wikigame indeed seem to prefer to navigate Wikipedia by following links of the underlying topological link network. This is not too surprising as the Wikigame per definition only allows users to click on available hyperlinks for trailing through the Wikipedia space. Additionally, we can see that the domain-specific similarity hypothesis is more plausible than both the universal uniform as well as the self-loop hypotheses. This corroborates the theories and assumptions of previous work [36, 45, 46] which observed that humans tend to follow semantically related concepts successively.

We also see that the universal uniform and the self-loop hypotheses are the least plausible hypotheses at interest. Interestingly, for \( k = 1 \) the self-loop hypothesis exhibits larger evidence than the uniform prior which partly also demonstrates that self-loops are indeed an existing aspect of this dataset as also observed in previous work [35]. However, with larger \( k \) the evidence of the uniform hypothesis surpasses the self-loop hypothesis which may be explained by the fact that we weight the informative part (i.e., only self-loops) too strongly while we ignore all other possible transitions.

Yelp. We depict the results for comparing the hypotheses at interest for our Yelp dataset (business reviews) in Figure 5(b). A first observation is that our approach indicates the domain-specific geographic hypothesis as the most plausible one. Hence, humans indeed seem to prefer to successively review businesses that are geographically close to each other on Yelp as captured by our dataset.

Contrary, the other domain-specific similarity hypothesis is less evident compared to the uniform hypothesis which can be seen as a baseline. Consequently, from this exemplary analysis, we cannot assume that humans prefer to consecutively review the same businesses based on similar categoric descriptors, at least not based on the similarity of categorical descriptors given on Yelp. Finally, the self-loop hypothesis is indicated as the least plausible one which indicates that humans at maximum very seldom review the same business twice in a row in our dataset.

Last.fm. Finally, the Last.fm results depicted in Figure 5(c) highlight that the similarity hypothesis, expressing our prior belief that users consecutively listen to songs that stem from the same artist, is the most plausible one. This is visible as the evidence values are larger for all \( k > 0 \) compared to the other hypotheses of interest.

Again, we observe that humans do not seem to prefer to listen to the same song over and over again (self-loop hypothesis) in our dataset. Also, for this example data, the track date hypothesis is less plausible compared to the universal uniform hypothesis.

5. DISCUSSION

The HypTrails approach represents an intuitive way of comparing hypotheses about human trails as we have demonstrated on synthetic as well as empirical data. However, there are some aspects— as partly exhibited throughout our experiments—that one should consider when applying HypTrails; we discuss them next.

Expressing hypotheses. When expressing hypotheses as matrices, choices have to be made regarding several factors such as (i) which transitions to believe in (e.g., about setting the diagonal) (ii) how to calculate values for representing a hypothesis (e.g., haversine distance for geographic closeness), (iii) how to normalize or (iv) availability of information (e.g., API restrictions). While several ways of doing this are conceivable, our approach does not constraint the researchers’ choice in this regard. If in doubt, our advice is to express the uncertainty through another hypothesis (which reduces the problem) or through a set of other hypotheses (which focus on different representations). For example, in this article we first investigate the plausibility of a universal self-loop hypothesis compared to a uniform hypothesis before making a choice about the diagonal of other hypotheses. We find that for navigational trails (Wikigame), self-loops seem to play a role at least occasionally (cf. Figure 5(a)) which is why we also set the diagonals in other hypotheses to larger values than zero, while for both Yelp (cf. Figure 5(b)) as well as Last.fm (cf. Figure 5(c)) we cannot observe such behavior. Another choice to make is whether one wants to express hypotheses in a symmetric or asymmetric way—e.g., it might be useful to believe that transitioning from state A to state B is more relevant than from B to A. Following our advice, we would express a symmetric and asymmetric version of the hypothesis and compare them.

Behavior of hypothesis weighting factor \( k \). Throughout our experimental results (see Figure 4 and Figure 5), we frequently observe that the evidence is falling with larger \( k \). As pointed out in this article, the evidence is a weighted likelihood average and is largest if both the prior as well as the likelihood concentrate on the same parameter value regions. The larger we choose \( k \), the larger we set the hyperparameters of the Dirichlet distributions and the more they get concentrated. Thus, only a few specific parameter configurations (single draw from the Dirichlet distribution) receive higher prior probabilities while many others receive low ones. As we cannot expect our hypotheses to concentrate on the exact same areas as the likelihoods as we did for our empirically aligned toy example in Figure 2(b), we sometimes see falling evidences with larger \( k \) as we reduce the scope of the prior. Even though we always want to compare hypotheses with each other for the same values of \( k \), comparing marginal likelihoods for different weighting factors \( k \) might give further insights. For example, the flat prior for \( k = 0 \) might be seen as a further baseline to compare to. Also, the behavior of the evidence for single hypotheses with varying \( k \) can give insights into the strength we can assign to a given hypothesis.

Memoryless Markov chain property. Currently, HypTrails is memoryless, meaning that the next state only depends on the current one. Previous work has been contradictory in their statements about memory effects of human trails on the Web (see e.g., [14, 35]). While first order models have mostly been shown to be appropriate, it may be useful to extend HypTrails to also support memory effects in the future. This would mean that it would allow us to not only analyze hypotheses about how the current state influences the next one, but also how past ones (potentially) exert influence.

Future work. In this work, we have suggested first ways of expressing hypotheses as well as eliciting Dirichlet priors. Future work should further refine these methods as well as study the sensitivity and interpretability of expressed hypotheses and elicited priors with respect to HypTrails; e.g., by focusing on how to normalize hypothesis matrices (potentially row-based) and refining methods for eliciting priors (distribution of chips).

Furthermore, while we have showcased a certain variety of datasets and hypotheses that can be analyzed with HypTrails, we would like to encourage researchers to see these examples only as a stepping
We highlight some exemplary findings next. Furthermore, one can also think about visual interfaces on websites; the importance of semantics between ample, in Figure 5(a) we show that both the structural as well as rationalities upon which humans base their navigational choices.

Hypotheses about memoryless transitions, the Markov chain model [17, 23, 31, 35, 43]). Given these advantages, and our interest in hypotheses about human trails on the Web. This work was partially funded by the DFG German Science Fund research project “PoSTs II” and by the FWF Austrian Science Fund research project “Navigability of Decentralized Information Networks”. Furthermore, we want to thank Alex Clemesha for giving us access to the Wikigame data as well as Daniel Lamprecht and Paul Laufer for valuable input.