ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/372443502

Tree-Test: an association test for observations on a directed tree

Thesis · September 2019

DOI: 10.13140/RG.2.2.36059.75047

CITATIONS
0

READS
26

3 authors, including:

Gal Novich
Amazon
2 PUBLICATIONS **2** CITATIONS

SEE PROFILE

All content following this page was uploaded by Gal Novich on 19 July 2023.

The user has requested enhancement of the downloaded file.

Tree-Test: an association test for observations on a directed tree

Gal Novich

Tree-Test: an association test for observations on a directed tree

Research Thesis

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Gal Novich

Submitted to the Senate

of the Technion — Israel Institute of Technology

Elul 5779 Haifa September 2019

The research thesis was done under the joint supervision of Prof. Roy Kishony of the Department of Computer Science and the Department of Biology, and Prof. Zohar Yakhini of the Department of Computer Science.

Acknowledgments

I'd like to thank Zohar Yakhini, and the Yakhini research group. Zohar was my first introduction to research, and what it is to research in the Computer Science faculty. The research group, which focused on the computational side of bioinformatics always kept me connected to my roots as a mathematician, data scientist, and machine learning researcher.

I'd like to thank Roy Kishony, and the Kishony lab. It has been my first introduction day to day life of a scientist and it is a great home of interdisciplinary research. Along with giving me the space to explore all kinds of hypothesis - Roy's guidance has been instrumental to my personal growth - transitioning from a theoretical mathematician to a skeptic researcher, learning how to set up a documented experiment - in code.

I'd like to give a special mention to Idan Yelin - which I had the pleasure to work alongside with physically and on the lab research projects. Not only is he a brilliant researcher, but a great co-worker and conversation partner. I'd like to mention the late Dr. Ning Yin. For the time that I had the pleasure of knowing him - I was blown away by this amazing mind, professionalism, his courtesy and sincere positive influence on all who were around him. He is a lasting positive influence on me. I only wish I knew him a whole lot more.

The generous financial help of the Technion is gratefully acknowledged.

Contents

List of Figures

Abstract

In classical statistics, when evaluating an association between two variables, independent observations are collected and statistical tests such as Fisher Exact are commonly used. However, for many real-world applications, the assumption of sample independence is quite erroneous. For example, in the field of population genetics - it is known that any collected observations are not independent, as they all share a common ancestor.

To answer the possible confounders that arise from sample evolutional dependence - geneticists have developed tree-based statistics. These tools aim to account for a hierarchical dependency structure of the samples dictated by the topological structure of their ancestry - a "family" tree. The current state of the art association tests use Monte-Carlo simulations to account for these dependency structures. However, the computational power needed to apply them is not negligible, making them unscalable for big-data analysis.

In our work, we introduce a generalized, simulation-free, analytic test that accounts for hierarchical sample dependency structures. We formulate our model assumptions, and compare our performance to the existing state of the art. Our method is widely applicable, as hierarchical sample dependency structures exist in many types of real-world data.
To showcase the strength and generality of our method, we present an analysis of big-data observational case studies of social media data sourced from YouTube and Wikipedia.

Chapter I

Introduction

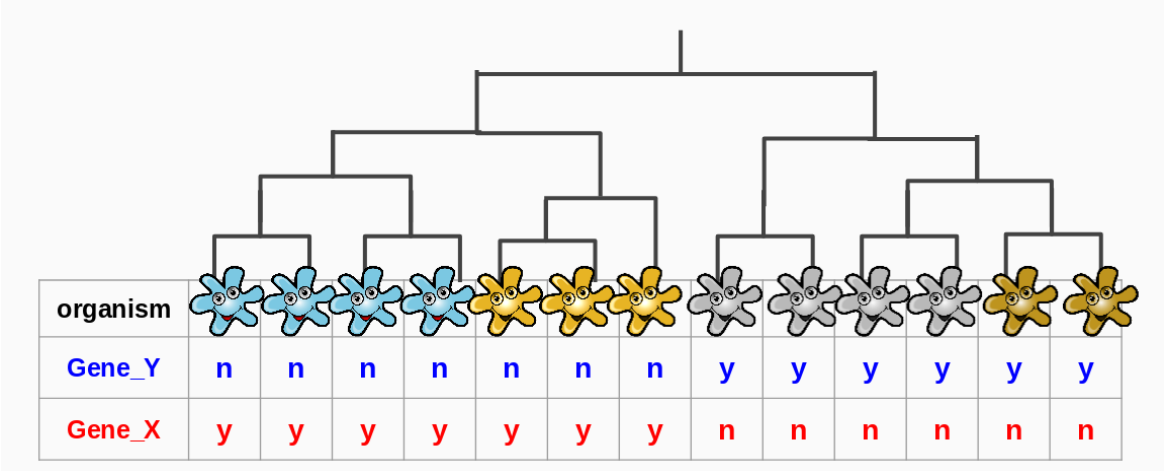
1. Problem setting

Suppose we would like to measure the association between two binary phenomena. For this, the standard statistical tool of evaluation would be the Fisher Exact Test [1], or it's approximation variant Chi^2-test [2]. These tests are used to try and reject the null-hypothesis of equally strong negative/positive correlation, under the assumption of independent sampling of our observations.

However, the assumption of independent sampling is not always viable. For example: In genetics, the theory of evolution recognizes that all different organisms have a common ancestor. The meaning of this is that observations from different organisms are not only dependent on each other, but their dependence can be described by their evolutionary ancestry - "The Tree of Life". Hence, the dependency structure of the samples is commonly interpreted as a tree, a "phylogeny", with our observations serving as the leaves of said tree.

This dependency structure becomes a relevant concern in Genome-wide association studies (GWAS). These studies aim to evaluate the association of genomic elements, "genotypes", with some biological phenomena, a "phenotype". For example, a phenotype can be some hereditary disease, and a genotype can be a mutation (SNP). GWAS have become the tool of choice in the early 2000s [3, 4], leading to the identification of many trait-associated mutations [5].

When our observations are dependent, a Pandora's Box of problems opens up. To visualize this, a schematic figure is brought here in *Figure 1.1.1*. For example, a researcher sampled several bacteria and tested the correlation of existence between 2 genes. When assessing the observations without ancestral context, they seem highly correlated. However, when sketching their evolutionary dependence, one can see that the gene existence signal might be confounded by the ancestry group from which the samples originate.



1.1 Schematic - Hierarchical dependence of observations

This is an extreme example of "Large-scale" confounding. However the "Fine-scale" dependencies, which may represent sample clonality/duplication, are equally as dangerous. Reexamine *Figure 1.1.1* and suppose that on some occasions, two closely related samples are in fact clones of each other, rather than an independent sample. That alone will be a confounder drastically skewing the observed signals.

The phenomena of population structure and the utilization of phylogenetic techniques are not limited to biology alone. Concept inheritance was long observed in other fields outside of biology [6]. Several surveys outline phylogenetic analysis of languages and the phylogenetic analysis of cultural artefacts [7], [8]. For example, historical-ancestry is often researched in linguistics [9]–[11], methods in this field have been dubbed "Computational Linguistic Phylogeny"s [12]. Tëmkin and Eldredge used phylogenetic methods to study the history of certain musical instruments, claiming that cultural artefacts, like genes and languages, reflect their history [13].

The potential of statistical confounding brought by population structure has been explored in multiple variations and formulations in and out of statistics as well. Next, we will provide a general overview of the efforts done thus far.

2. Methods overview

a. Clustering methods

Population stratification - where there are inherit frequency differences between differing known subgroups in our study - can cause spurious associations in any study. Many methods have been utilized to deal with such phenomena, such as Cochran-Mantel-Haenszel correction [9, 10], Principal Components Analysis [16], and Dimensional Reduction techniques [17].

These methods assume some stratification of the samples, or group them into clusters them by k-means algorithm or one of its computational variants. With the groups recognized, the representation of each group can be accounted for in the analysis. Although these methods might partially account for large-scale ancestry differences, fine-scale sample dependence within each group is left unanswered by these tests alone.

Fine-scale differences, like clonality/duplication, are often accounted for by sampling methods. Related samples are identified by thresholding of hierarchical clustering or nearest-neighbor algorithms. Then, multiple tests of association are done by sampling, and their combined estimate yields the prediction.

Combinations of “Large-scale” and “Fine-scale” methods are possible, but are indicative of the sometimes recursive nature of these dependencies - when do we stop partitioning to groups? Moreover, the thresholding process of “Large” and “Fine” can be arbitrary and ad hoc. For this reason, approaches that account for all possible dependencies are often used in place.

b. Phylogenetic approaches

Phylogenetic trees allow for the detailed identification of relationships, not only at the level of population clusters, but also at the resolution of subpopulations and individual relationships [18]. They have been found to be helpful, sometimes crucial, to a comparative analysis.

Algorithms for the estimation of phylogenies from empirical data are a main research field in genetics, dubbed “Ancestral Reconstruction”. Classical algorithms such as Maximum Parsimony, and Maximum Likelihood [19]–[21] were devised for the joint estimation from empirical data of both the tree structure, and the state of the internal nodes of the deduced tree.

Bayesian approaches have also been introduced [22]–[24]. The task of identifying a possible tree topology alone can be done by standard Hierarchical clustering algorithms such as UPGMA, NJ, etc. [20, 21]. However, even with a known topology, the "Ancestral Reconstruction" problem is difficult on its own [27].

After the dependency structure has been defined, there are several methods to account for it the analysis of association. The popular examples for continuous traits are notably Phylogenetically Independent Contrasts [28] and its generalization - Phylogenetic Generalized Least Squares [29].

There, the phylogenetic tree is converted to a symmetric matrix, intended to represent expected variances and covariance of "leaf data". The incorporation of such variance–covariance matrices into standard linear models has been done in many variations since [30].

Markovian modeling for binary trait analysis have been introduced by Pagel [26, 27]. The modeling calculates the log-ratio statistic of fit difference between an independent model and a possibly dependent model. This test statistic is asymptotically distributed as a chi^2, with the null hypothesis of independence. One known drawback of the evaluation is that it falters when the phylogeny contains a small number of samples, or rates of transition are low.

Moreover, the computational requirements of fitting the models for every trait pair make it unviable for large-scale analysis. Direct combinatorial approaches that calculate the probability of seeing an association of binary variables exist [33]. However these suffer from tedious recursive calculations, and as such are not scalable.

Monte-Carlo simulations have been proposed in the 90's [29, 30]. These methods are designed to directly account for a given tree, simulating many traits along the topology to evaluate a null distribution of association. They are overall favored for their accurate and reliable performance, and have been commonly used up until today. Over the years many simulation architectures have been introduced, adding more possible parameters to be accounted for, or alternative metrics of association [31, 32].

TreeWas [38] is the current state-of-the-art tool for this approach. In short, given an empirical dataset and a target trait - it estimates the mutation rate distribution of the empirical dataset using maximum parsimony, and conducts many

simulations - giving a null hypothesis distribution of the expected association score with the target. It is very flexible - supporting several scoring methods, ancestral reconstruction methods, assumption models etc.

Although robust and accurate, Monte Carlo simulations still require considerable computational resources. The evaluation time with one phenotype is approximately linear in the number of individuals and simulated traits. As such, they are not scalable for “big data analysis” of pairwise correlations between many traits over big trees. TreeWas reports that for one phenotype and a tree constructed from ~12k observations, it would take approximately an hour to finish the evaluation for ~10k traits.

Chapter II

Tree-Test

1. Motivation

Our aim is to establish an efficient analytic framework that captures the information of the hierarchical dependency structure. We will be avoiding recursion or simulations, thus saving computational time. We must also strive to provide as good an estimate as the state-of-the-art simulation frameworks do.

Here, we will articulate the main observation the gives way to our formalization:
The estimation of dependence of two traits involves an estimation of their shared covariance. “Covariance” being in layman's terms - the amount in which the two traits change together. Suppose the hierarchical dependency of our observations was available to us, with complete ancestral reconstruction. Our observation may not be dependent, but the events of change along ancestry - are.

We aim to directly quantify the presence of covariance between variables. However, for the statistical evaluation of significance - some model must be assumed or estimated. We carefully choose a limited number of assumptions under which a Markovian model can be reasonably and efficiently estimated. We will start our formalization with simplistic assumptions, only to generalize upon them for our final terminology. To our knowledge, we are the first to introduce this type of formulation.

2. Definitions

We will now lay out the definitions for a hierarchical dependency structure, and the score we utilize to quantify the covariance. We will start by a simple formalism. We will later introduce a generalization for traits produced from a Markovian process. In this work - which aims to apply the model as a proof of concept - we will only be using/applying the formalism in *sections a-c*.

a. Covariance Score

Definition 2.1.0 [*Observation Tree*] Given traits *X,Y* and a set of observations $S = \{(x_i, y_i)\}_{i=1}^{m}$, *an Observation Tree over S is a directed tree with vertices exactly our set S, T=(S, E) .*

The Observation Tree can be generalized to an "Observation Forest", or any rooted directed graph for which the in-degree of all nodes is at most 1. W.l.o.g we will limit our definitions to a single connected component.

Definition 2.1.1 [*Edge Observation*] *Given an Observation Tree, the Edge Observation of* $e = (v_{par}, v_{des}) \in E$ *for trait X is defined as:* $x_e = (x_{par} - x_{des})$. *The Joint Edge Observation for traits X and Y is defined as:* $x_e y_e$

Notice the possible values of an edge observation are a symmetric set. The Joint Edge Observation represents the trajectory of the joint change in both *X* and *Y*. It is zero if at least one of the properties does not change over the edge, positive if they change with the same trajectory, and negative of opposite trajectories.

Definition 2.1.2 [*Covariance Score*] *Given an observation tree T=(S, E) over traits X and Y, the Covariance Score of X and Y under T is:*

$$(1)\ CovScore_T(X, Y) = \sum_{e \in E} x_e y_e$$

The covariance score quantifies how much did the two traits synchronously changed between our dependent observations together. If they haven't changed at all - the score would be zero. A positive score would indicate that the properties come and go together - mutual dependence, while a negative score would indicate mutual exclusion.

The covariance score will be used as our "test statistic". The following definitions will establish our statistical modeling and assumptions for the assignment of significance:

b. Random Edge Variable

Definition 2.2.0 [*Edge Random Variable*] *Let X be a binary trait, with a "symmetric change rate"* $p_X \in [0,1]$. *Given an Edge Observation* $e = (v_{par}, v_{des}) \in E$ *from an Observation Tree T. The Edge Random Variable* X_e *is defined as:*

$$P(X_e = 0) = 1 - p_X$$
$$P(X_e = 1) = P(X_e = -1) = \frac{1}{2} p_X$$

The assumption under this modeling is that $P(v_{par} = 0) = P(v_{par} = 1)$. The reason for this assumption will become clearer further on in our definitions.

For two independent binary traits X,Y, with change rates p_X,p_Y - we will denote the Joint Edge Random Variable for the Joint Edge Observation one as X_eY_e.
As X,Y are independent, the probability function of X_eY_e would be:

$$P(X_eY_e = 0) = 1 - p_Xp_Y$$
$$P(X_eY_e = 1) = P(X_eY_e = -1) = \frac{1}{2}p_Xp_Y$$

Definition 2.2.1 [*Covariance Score Random Variable*] *The score between X,Y under the edge Joint Edge Random Edge Variables is defined as:*

$$(2)\ Cov_T(X,Y) = \sum_{e \in E} X_eY_e$$

Notice the connection with formula (1). The observations have been supplanted by our above defined Edge Random Variables.

c. Tree-Test

Given an Observation Tree T of traits X,Y. One can try and reject the null hypothesis of X_e, Y_e independence, by calculating the two-tailed test:

$$p_{value} = P(|CovScore_T(X,Y)| \leq |Cov_T(X,Y)|)$$

The calculation of the distribution of the Covariance Score Random Variable can be done directly under the null hypothesis of X,Y independence. One can use $|E|$convolution steps of X_eY_e, yielding the exact distribution, or by the following approximation given by the Central Limit Theorem:

$$(3)\ \lim_{|E|\to\infty}\sqrt{|E|}Cov_T(X,Y) = \lim_{|E|\to\infty}\sqrt{|E|}\sum_{e \in E} X_eY_e =$$
$$= N(0, Var(X_eY_e)) = N(0, p_Xp_Y)$$

To summarize our assumptions, given two traits X,Y for which:

- Change events are i.i.d
- Change rates are constant
- Loss and gain of a trait are equiprobable

One can try and reject null-hypothesis of X,Y independence by collecting an observation tree of hierarchical dependencies and use branch observations in a Tree-Test. The change rates can be estimated in a boot-strapped fashion from the data by Ancestral Reconstruction methods.

The model is based on the evidence of change between dependent samples, and therefore isn't always applicable. For example, when considering a forest of singletons - all observations would be independent, and no edge observations can be derived.

d. Markov chain formulation

Markov chains are canonically defined as a series of variables $\{X_i\}_{i=1}^{m}$, with the markov property: the probability of moving to the next state depends only on the present state. We formulate a time series analysis estimate of the random variable defined by the difference in an arbitrary step i: X_{i+1}-X_i.

Let us assume a trait X is produced by a discrete time-homogeneous Markov process of real-value states in state vector s, a transition matrix P, which has an equilibrium distribution over the states $\boldsymbol{\pi}$.

Definition 2.4.0 [*Markov Edge Random Variable*]
The "Markov Edge Random Variable X_σ" corresponding to X is defined by:

$$(4)\ P(X_\sigma = x) = \sum_{k,l:\ s_l - s_k = x} p_{kl}\pi_k$$

This is an approximation for the expression "X_{i+1}-X_i ", under the assumption of $P(X_i = s_j) = \pi_j$. We claim this assumption is statistically fair: The probability of being in a certain state in some arbitrary time after many transitions, would be the stationary distribution. Therefore, if many edge observations have been recorded - many transitions have indeed occurred. In other words - the significance is directly dependent on the number of observations. Note that for the general transition matrix P an equilibrium distribution is not guaranteed.

Tree-Test for Markov chains:
Let there be two traits as described above X and Y, with corresponding ($P^{(X)}$, $s^{(X)}$, $\boldsymbol{\pi}^{(X)}$), ($P^{(Y)}$, $s^{(Y)}$, $\boldsymbol{\pi}^{(Y)}$). Let X_σ, Y_σ be the corresponding Markov Edge Variables.

Given an Observation Tree T of traits X,Y. Under the assumption of independence of the edge observations, one can try and reject the null hypothesis of X, Y independence by calculating the two-tailed test:

$$(5)\ P(|CovScore_T(X,Y)| \leq |\sum_{e \in E} X_\sigma Y_\sigma\ |)$$

The modeling in *sections a-c* is a Markov processes for two binary states with the transition matrix - $P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$, for some parameter *p*. Although we do not explore it in the scope of this work - we know that for a transition matrix $P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$ the stationary distribution is - $\pi = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$. Therefore, binary loss-gain rate modeling is equally as feasible under our framework - although more data will be needed for the estimation of this 2 parameter model.

The generalization of *section d* frames Tree-Test as the evaluation of synchronous behavior between two seemingly independent discrete-time Markov processes over real-valued/ordinal states. Although not analyzed in this work, ordinal traits are interesting not only for possible applications purposes - but as a bridge for the possible extension of the model to continuous traits in future work.

3. Model evaluation via Simulations

Now that we have established our model theoretically, we would like to validate it empirically. To do so - we will compare it to the Monte Carlo state-of-the-art framework: TreeWas. We compare the models performance under the assumptions defined above in *sections a-c.* We expect TreeWas to have optimal performance, as its generalist, robust, approach is more than capable of handling our limited setting. What we aim to see is an equivalent performance under our model assumptions. These would establish the analytical approach and make the need for simulations redundant.

a. Simulation architecture

For a fair comparison, we have many elements in our system to be controlled for i.e a tree topology, change rate, target trait etc. We control all these factors by running many independent simulations. The process of creating one simulation is the following:

1. A random tree topology T of size M is generated.
2. A change rate p is set and $N+1$ traits are generated, with a target trait Y.

For each simulation, we will be comparing the performance of significance evaluation to a "ground truth significance":

3. We calculate the Covariance Score for each trait and a rank order p-value is assigned. This set of p-values is denoted as the vector v_{truth} of size N.

Our Tree-Test at this point would be able to give his own evaluation. However, this would not be a fair comparison with the TreeWas pipeline. We evaluate our performance on equal grounds: where only the leaf information is available to us.

4. For each trait, we use Maximum Parsimony - yielding an estimated change rate set $\{p_i\}_{i=1}^{N+1}$, and ancestral states. The target trait Y is assumed to be reconstructed perfectly (w.l.o.g, as an error will harm all estimates).
5. The covariance score is recalculated for each trait by the ancestral states.
6. A Tree-test p-value, $v_{\text{tree-test}}$, is assigned on the reconstructed traits.

At this point TreeWas will need its own nested round of simulations:

7. TreeWas constructs a rate distribution h, gathered by the rate set $\{p_i\}_{i=1}^{N+1}$.

8. TreeWas constructs a null hypothesis distribution of Covariance Scores H, as gathered by an additional set of simulated traits, each with a generated with change rate sampled from a rate distribution h.
9. TreeWas assigns significance to the traits by H, yielding $v_{TreeWas}$.

b. Results

We run our simulations ten times using M=50k, and the parameter grid:

- $N \in \{100, 250, 500, 1000, 2000\}$
- $p \in \{\frac{10}{2N}, \frac{25}{2N}, \frac{50}{2N}, \frac{75}{2N}, \frac{100}{2N}\}$

Our performance is evaluated by our success in estimating the ground truth p-value vector. The task of comparing two p-value vectors can be approached in several ways, and there is no strict consensus. We will weigh our p-value vectors, as our concern is to avoid false negatives - spotting the rare events correctly. In all figures, we weigh our vector by the ground truth significance. *Figure 2.3.1* visualizes the performance under a fixed population size of *N=500*, with increasing change rate *p* on the x axis. In *2.3.1* we set the pairwise Pearson correlation as the y axis, and in *Figure 2.3.2* - the pairwise log-distance.

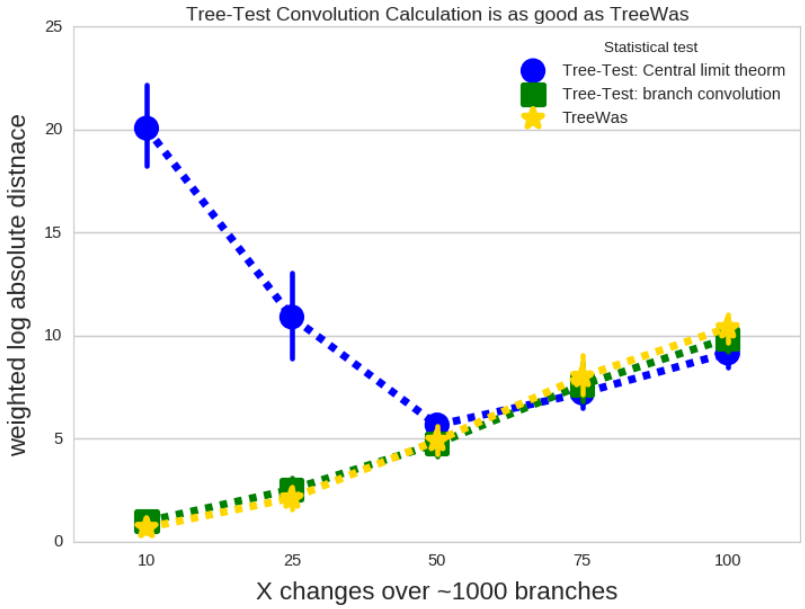
Figure 2.3.2 mainly indicates to us that our tests are empirically consistent with the ground truth. Let's explore *Figure 2.3.1*, starting with TreeWas. One can see that as the change rate increases - so does the weighted error. This can generally expected, as higher number of change events are harder to ancestrally reconstruct. False ancestral reconstruction directly translate in an estimation error - which is the error we evidently see.

Although successful in high transition rates, the central limit theorem flavor of the tree-test suffers on low transition rates. This can be seen in both *Figure 2.3.1*-2. The behavior is consistent with the asymptotic evaluation of Markov modeling by Pagel [26] mentioned above. In their original paper, they noted the same weaknesses exhibited in our simulations. The CTL Tree-Test and the models by Pagel are indeed equivalent in the sense that they share the same asymptotic assumptions and closely related Markov modeling.

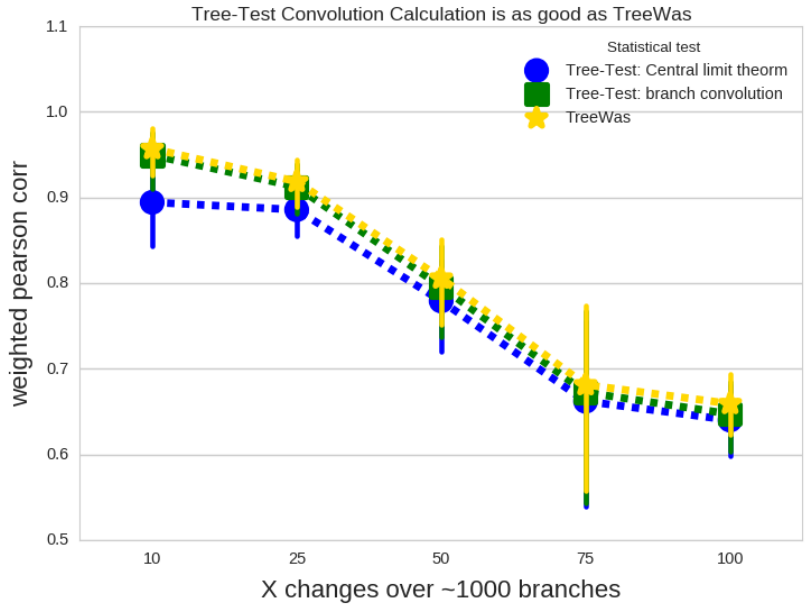
We would have liked to provide a direct empirical comparison with BayesTraits - the software by Pagel. However, the computational time required to fit Markov models per evaluation would have been un-feasible for large M (with an estimated 15 min just for M=100). For this reason we are providing a limited

small scale assessment in *Appendix 1*. There, we incorporate BayesTraits into our pipeline, and discuss the possible strengths and weaknesses of the method.

We are happy to report that our convolution flavor Tree-Test has been able to compensate for the low change rate error, as indicated by *Figure 2.3.1*-2. The combined time taken by both Tree-test methods for the estimation of M=50k traits, including ancestral reconstruction, score assignment and pval calculation was under 3 seconds. Most of the time was spent on the reconstruction and scoring themselves, while computation time differences of CLT and convolution Tree-Test were negligible. As the convolution test has empirically outperformed the CLT Tree-test - we will be focusing on the former for the rest of this work.

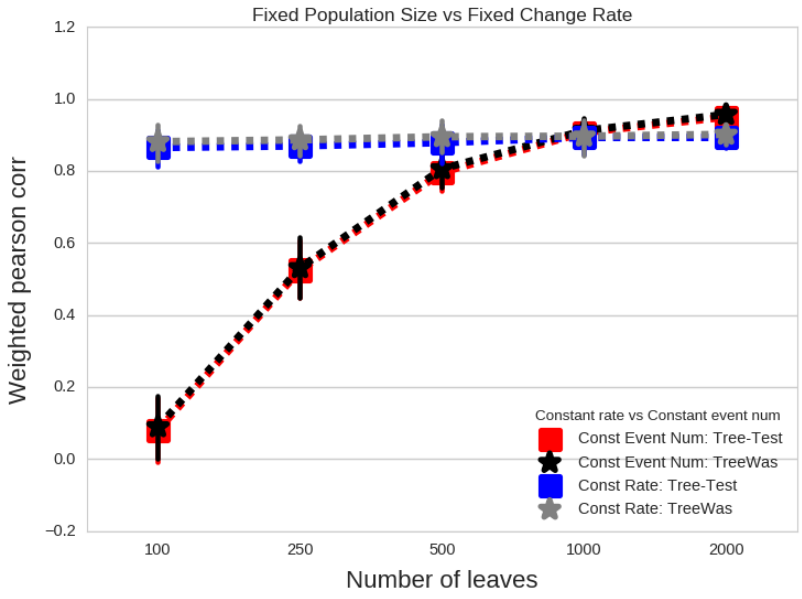


3.1 Tree-Test vs TreeWas: distance



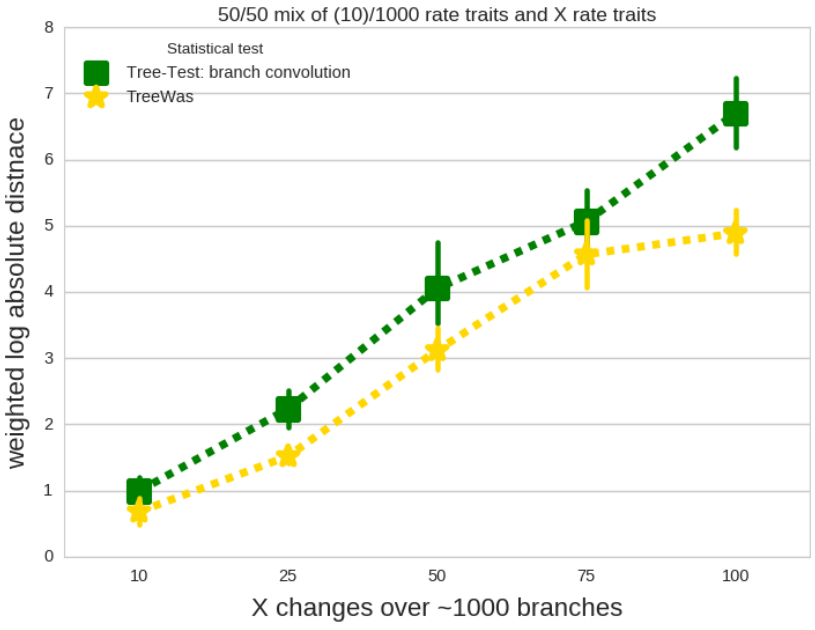
3.2 Tree-Test vs TreeWas: correlation

We are checking the validity of our modeling under an increasing change rate. In the next experiment we want to make sure the size of the population simulated does not affect our results. For this we have simulated the same rate of change on increasing population size, and contrasted it with a fixed number of events. The results are shown in *Figure 2*.3.3. We see that the constant rate has equivalent performance - as anticipated. As the size of the population increases and the number of events stay the same - the change rate decreases - yielding better performance, as previously seen in *Figure 2.3.1*-2.

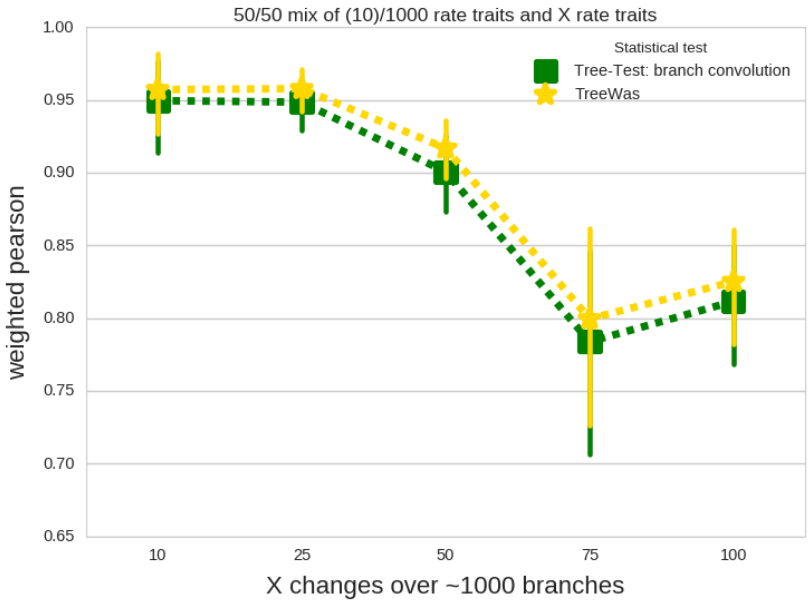


3.3 Size vs rate sanity check

For our last simulation, we tried to introduce a confounder that would disrupt the assumptions of our modeling. The main one being - a constant rate of change over all traits. To disrupt this we have given as input a trait pool sampled equality from 2 subpopulations with differing change rates. One constant and low, the other increasing for each independent simulation. This difference will disrupt the rank-order ground truth p-value in a way which Tree-Test would not be able to account for. TreeWas would though, as it will estimate the transition rate distribution - recognizing the bi-modality of the trait population. The results are presented below in *Figure 2.3.4-5*. In *Figure 2.3.4*, as seen in *Figure 2.3.1*, as the change rate increases so does the error. However, unlike *Figure 2.3.1* the equivalent performance of tree test has not endured.
Nevertheless, we learn from *Figure 2.3.5* that when considering correlation - these differences may not be substantial. One can rationalize this by recognizing that separately - the 2 trait subpopulations will be correctly ordered for significance by Tree-Test. The only hindrance on the correlation with the ground truth then would be the resulting interleaving of these 2 ordered trait subpopulations. In future work, it would be interesting to quantify the expected error of such trait population mixtures.



3.4 mixed trait population test: distance



3.5 Mixed trait population test: correlation

Chapter III

Case studies

We have established that under the right conditions, our statistical test performs as well as the state-of the art of the field. We now turn to present possible applications. To illustrate the effectiveness of Tree-Test - large-scale analysis of pairwise comparisons will be performed. The analysis examples presented would have been hard to perform in the TreeWas framework, as it would have needed a separate simulation run for each trait, making it unviable.

To showcase the strength and generality of our method - we will establish that the Tree-Test framework is applicable in situations spanning way beyond its Genetics origins.

1. Case Study I: Observable hierarchical bias.

a. Motivation

When planning the possible applications to our test, we could not ignore the following thought: "Wouldn't it be better if we knew for certain the trait values in the internal nodes, and not estimate them?". In what real-world situations is the ancestral reconstruction redundant?

Many examples reveal themselves when considering human hierarchies. In medicine, a family looking for generation-wide genetic tests has a known full ancestry. Companies have employee hierarchies. In social media: Group's have an admin/founder, who invites friends, which in turn invite more friends.

In our first case study, we consider the underlying hierarchical structure of Wikipedia articles. In Wikipedia, "Category Pages" store many articles, and other subcategories under them. For example, the "Category:Physics" page has 26 subcategories: "Concepts in physics", "Physics Literature" etc. and 55 pages directly under it: "Action-angle coordinates", "Silicon nanowire" etc.

Moreover, many category pages have a unique "main article" documentation for them. For example, "Category:Physics"'s main article is "Physics". When supplementing the categories with the main articles, a hierarchy of Wikipedia articles is formed. Here, nodes are the articles, and directed edges represent the category structure.

With this in hand, one can use Tree-Test to evaluate co-variation of words. This type of evaluation would give us the power of estimation of "the language of the field". This analysis provides a refreshing new view on computational semantics. See *Appendix 1* for a more visualized example.

b. Dataset Assembly

Here, we will outline our pipeline for the construction of the Wikipedia dataset. While the goal of Wikipedia's categorical features is to provide a hierarchy structure of all information, it is not a perfect tree. Many articles have several categories associated with them, and even some outliers of cycles have been documented. Nevertheless, trees can be derived from it with ease:

1. Select a starting Category Page, and set it as the root node.
2. Run a Limited-bandwidth BFS - given a category page:
    a. Randomly select a subset of up to c yet unseen subcategories, and a yet unseen articles. Set them as descendants.
    b. Add subcategories to queue of expandable nodes.
    c. The articles are to be interpreted as leafs.
    d. The LBFS halts after collecting N articles.

3. Most Category pages have a “main article” page associated with them. When available, retrieve and supplant them as inner nodes.
4. Tokenize words in articles, and keep those that appear in at least *n*.
5. Traverse post-order: If an inner node has no “main article” - use the Maximum Parsimony, setting the most common state among the immediate descendants. Make sure to cut “subcategory leafs” that haven't been expanded upon.

The construction of the database was made via the MediaWiki API. In our proof-of-concept setup, we selected to explore the linguistics of Mathematics. We started from the root “Category:Mathematics”, and ran the above scheme for a maximum branching factor (c)=5, maximum leaf per internal node (a)=5, number of articles collected before stop (N)=1000, and minimal occurrence threshold (n)=3. The resulting dataset contains over 1000 articles with 5683 binary token-traits. We provide a visualization of a subtree of the resulting dataset *Appendix 1*.

c. Results

Here, we showcase some analysis examples of our dataset. We start by selecting a target trait - a mathematical term. Then, all other traits are evaluated by our Tree-Test and a Chi^2-test. The Chi^2-test is chosen as the normal approach a researcher would have used if he were to scrap all the pages and treat them as independent.

To have some ground truth estimation of the strength of relationship between two words, we utilize a technique from the field of Natural Language processing in the form of WordNet and the Wu-Palmer metric [35, 36]. WordNets are directed rooted graphs that aim to map the semantic association of words. WordNets have been manually constructed in the 80’s and 90’s by psycholinguistics. Even today, after some automation have been introduced, manual curation and optimization is heavily used in their assembly. WordNet, still holding the title of most widely used lexical resources in natural language processing and has not been updated significantly since 2006 [41].

Given 2 words (nodes) on a WordNet, the Wu-Palmer metric is defined as the path which maximizes the following expression:

$$sim_{WUP} = max\left[\frac{2 * depth\big(LCA(a,b)\big)}{length(a,b) + 2 * depth\big(LCA(a,b)\big)}\right]$$

In words: the minimal weighted path between them on the WordNet, normalized by the depth of the least common ancestor (also known as "Subsumer").

i. "Sin"

Here we provide an example analysis of the trigonometry term "sin". We will first illustrate a general view of the results with a scatter map. On the left of *Figure 3.1.1*, the Chi^2 test is summarized by a log-log scatter plot of words. The words are plotted by the score of absolute odds-ratio over the p-value, with color that indicates the Wu-Palmer Score. The right side of *Figure 3.1.1* does the same for Tree-Test, using the covariance score.

In both figures, thresholds of significance are plotted as dashed lines. The vertical line indicates the quantile of 5% top scores. The horizontal line represents the minimum between the 5% top p-values and the threshold for which samples pass 5e-2 p-value after a Bonferroni correction over <u>all</u> word pairs in the analysis.

After examining the scatter plot we were delighted to see a significant Pearson correlation between the covariance score and the Wu-Palmer metric, with a 0.2 slope with 4.3e-56 p-value. This is comparatively much better than the odds-ratio score - with 0.034 slope and a p-value of 8e-3. Visually, this can be estimated by seeing greener dots as the score rises.

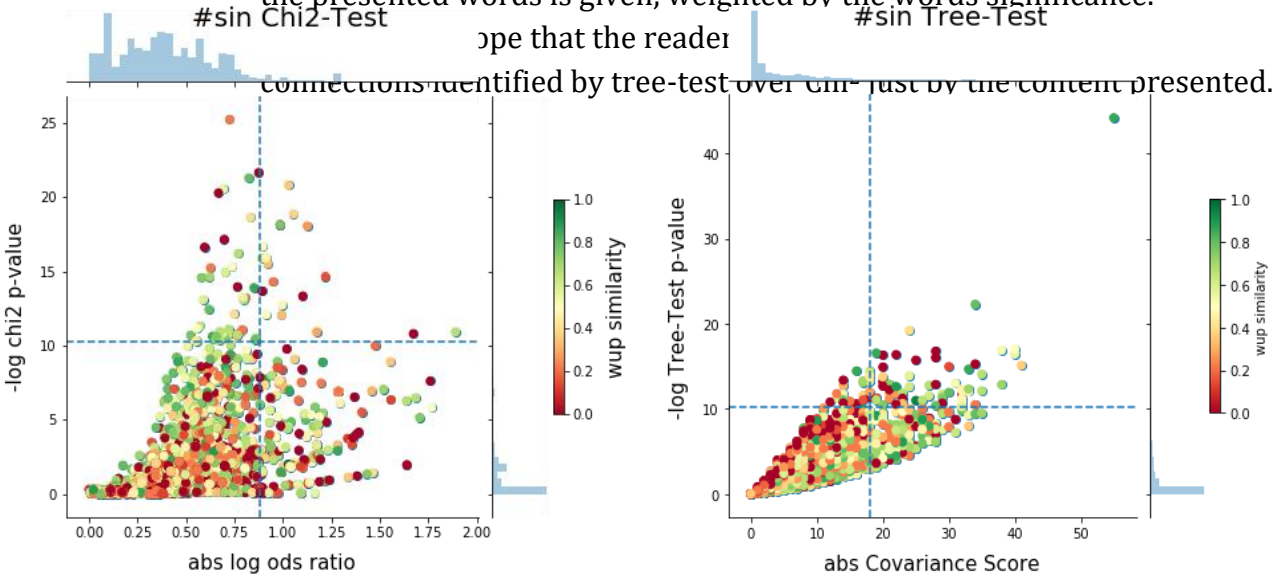
The main claim here is that Tree-Test is better suited for the identification of synonyms, and filed semantics. It is not that Chi^2 test has any inherent faults in it - it only estimates with different goals in mind. From Chi^2 we learn that words are seen more than to be expected with "sin". From Tree-Test we learn that words come and go together more than to be expected with "sin".

We are interested in what words are the significant top scoring ones. To visualize them, we use the word-cloud technique. We plot only the words that pass our thresholds of significance, score wise and p-value wise. In *Figure 3.1.2*, these words are plotted for Chi^2 test and Tree-test

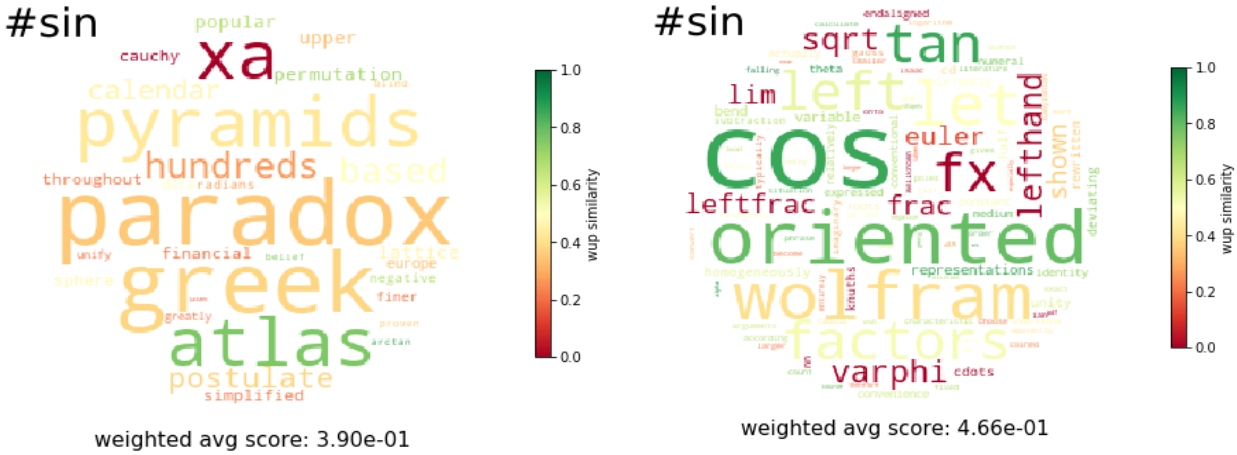
respectfully. The word size is inverse proportional to the negative log p-value, while the color is the Wu-Palmer metric - which stays consistent with the scatter plot.

To compare between the two tests, an average of the Wu-Palmer metric of the presented words is given, weighted by the words significance.

connections identified by tree-test over chi- just by the content presented.



1.1 "Sin" - Scatter plots



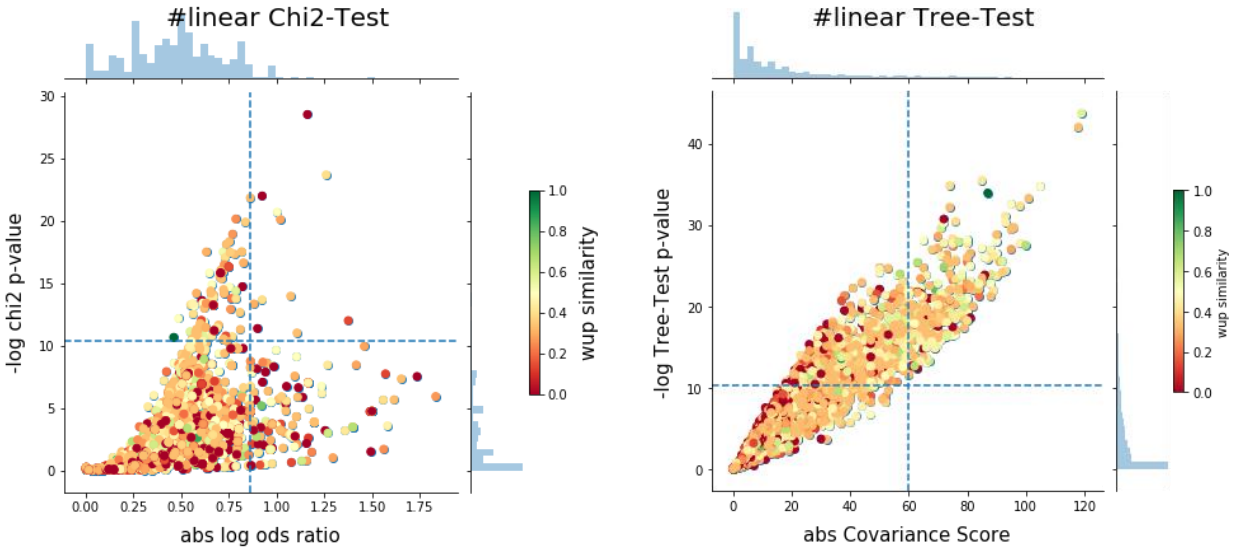
weighted avg score: 3.90e-01

weighted avg score: 4.66e-01

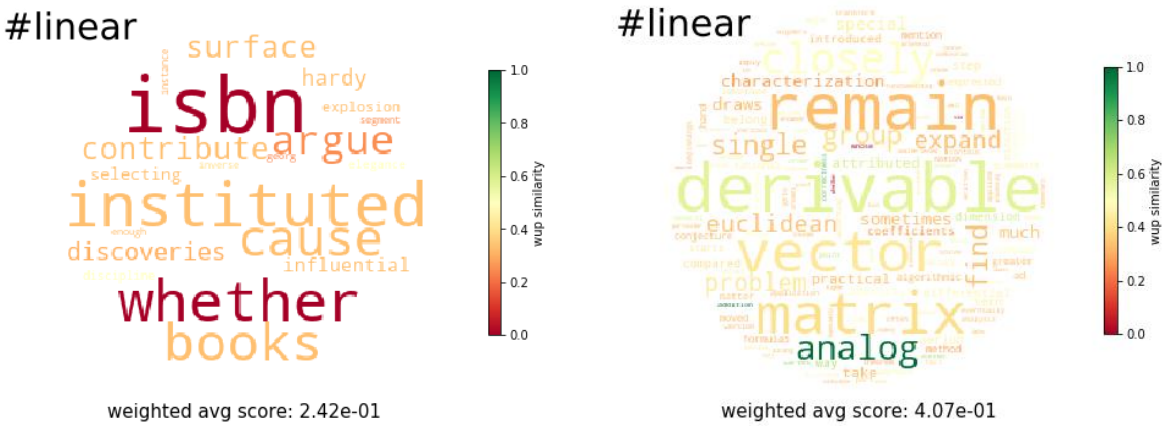
1.1 "Sin" - WordClouds

ii. “Linear”

To further substance our claim, we provide a second analysis in the same vein of the term “linear”, presented below in *Figures 3.1.3-4.* Here, the Pearson correlation between the covariance score and Wu-Palmer is even stronger, with a documented 0.32 slope with 2.35e-140 p-value, versus the odds-ratio's 0.023 slope and p-value of 8e-2.



1.3 “Linear” - Scatter plots



1.2 “Linear” - WordClouds

2. Case Study II: Un-Observable hierarchical bias

a. Motivation

We now turn to an application example where a suspicion for hierarchical dependencies exist, however clear ancestry information is unavailable. Good examples of this phenomena are naturally found in biology - were genes are inherited by evolution. In these cases, observed organisms are known to have a common ancestor, but in most cases - their ancestor is unavailable to us, and can only be estimated.

Many fields outside of biology might utilize such a perspective. Social sciences regularly try to account for community structures such as, town, province, country, etc. Analysis of media, such as films, should take into account genres, subgenres and "auteurs" with signature style. A retailer, performing an analysis of its product selection has to consider subtypes of items, and brand offerings.

We continue our social-media oriented motivation in a dataset of trending YouTube videos. YouTube hosts video content of endless variety - news, entertainment, sports etc. Each video has a wealth of data: title, category, uploader, uploader social hashtags, etc. (example: Tracking the path of Hurricane Dorian, News & Politics, CBS News, #Dorian #Miami #News)

Works have been done on the interpretation of the "hidden language of internet" [41]. We would like to use Tree-Test for the analysis of trending, topical, data. As before, we will aim to evaluate the covariance of terms, here being the hashtags given by an uploader to his video.

b. Dataset Assembly

The dataset was retrieved from the following online source [42]. It contains 6k videos "trending" in the US collected from 2001 to 2012. To evaluate their dependence, we utilize UPGMA hierarchical clustering algorithm. The clustering was based on a one hot encoding of category and uploader, and so the hierarchy is "independent" from the hashtag data.

As mentioned, the binary traits evaluated here are the occurrence of hashtag expressions. We tokenize every word that appears in uploader tags, completing the observed data on the tips of our dependency tree. Then, we ancestrally reconstruct token states via Maximum Parsimony. Unlike the previous chapter,

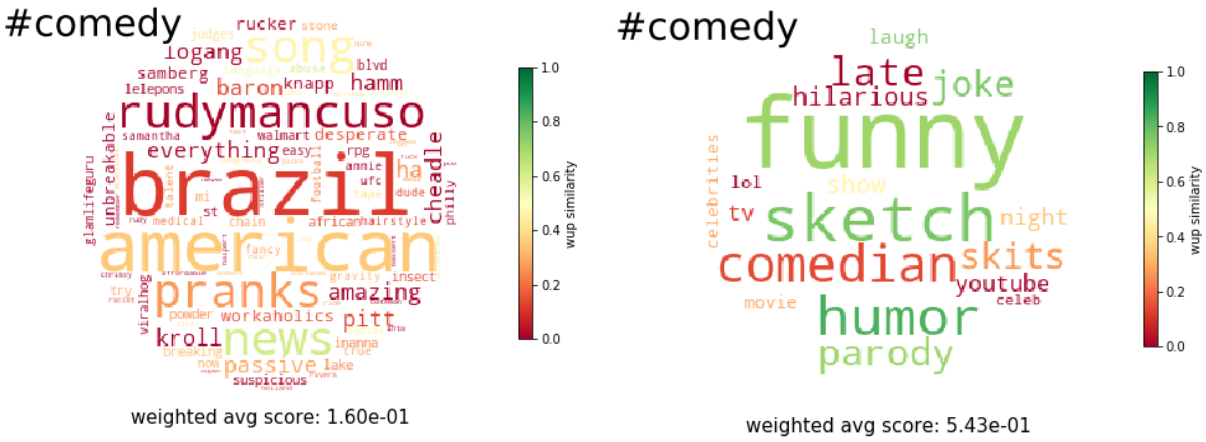
we perform the Chi2-test on the tip data only - as a researcher would have naivly done after collecting his observations. We also calculate Tree-test, and Wu-Palmer word similarity metric, as we did previously. In total, 6244 trending videos were used, and 6118 word-tokens were evaluated.

c. Results

We showcase our results, using 2 examples with the same visualization method described above in the results section of Case Study I.

i. "Comedy"

In *Figure 3.2.1* the reader can see a great Wu-Palmer statistic difference, and a general sense of accuracy given by the word selection of Tree-Test. We are most excited by the inclusion of the term "lol" in the significant high scoring terms. "Lol" - "laughing out loud" is a classic "internet speech" acronym, directly related to comedy and very appropriately used to express hilarity.



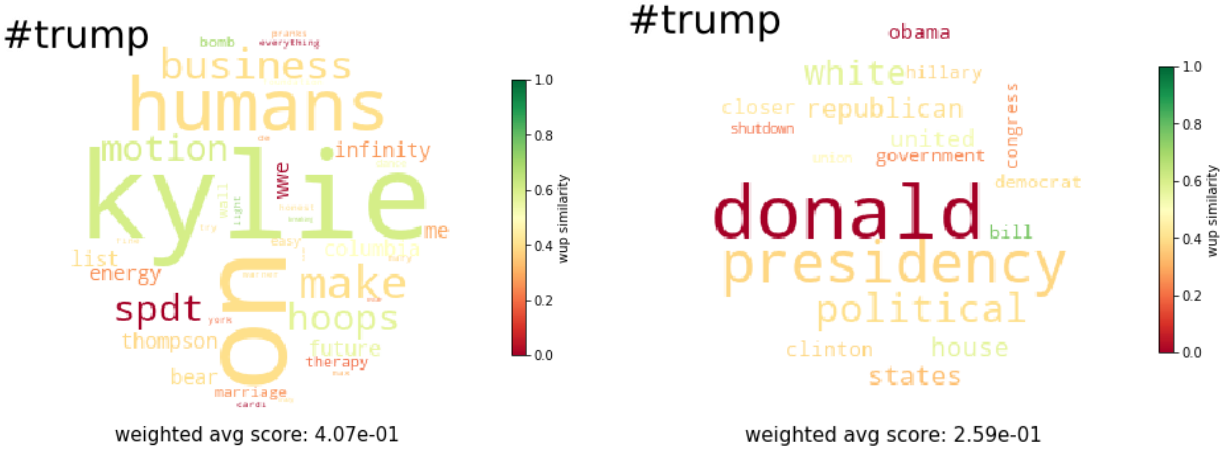
2.1 "Comedy" - WordClouds

ii. "Trump"

We conclude with an example for when the Wu-Palmer metric derived from WordNet does not sufficiently reflect the ground truth. Topical terms, and words not necessarily grounded in the English language, will not be well represented in WordNet corpus. To showcase this, we chose the

contemporary term of president Trump. The analysis results are shown in *Figure 3.2.2*.

After researching the results given by chi^2 test, we have concluded that videos reporting on topical news, also indeed happen to report on several celebrities named “Kylie”, such as “Kylie Jenner”. There are many interesting phenomena worth mentioning in the Tree-Test Results: Trump is associated with the widely reported “government shutdown”. Political figures, not represented in WordNet corpus, such as “Obama” and “Clinton” are prominent in the results. A greater significance has been assigned to “republican” over “democrat”. Etc.



2.2 “Trump” - WordClouds

Chapter IV

Conclusion and discussion of future work

1. Conclusions

In this work, we have introduced an analytical method for the evaluation of binary and ordinal trait associations by hierarchically dependent observations. The method utilizes Ancestral Reconstruction methods to boot-strap the modeling of ergodic, time-homogeneous Markov process for each trait. With the modeling efficiently estimated, direct quantification and statistical evaluation of their covariance is computationally cheap by a finite number of convolutions.

We compared our test to the state-of-the-art simulation techniques and deduced their equivalent performance under our assumptions. As our model is highly efficient, it enables us to perform truly Big-Data analysis to identify many trait-pairs. We then applied our work to wide-scale analysis of binary traits sources from media platforms of Wikipedia and YouTube. The pairwise comparisons estimated by our approach can be used to estimate topical and topic-professional semantic maps.

One of the goals of this work was to apply biologically motivated algorithms in non-standard applications. For this reason, we consistently proposed possible projects along this work for the reader's motivation. We hope our wide-scope approach would promote awareness to hierarchical dependence in statistical tests. Moreover, the efficiency of our method hopefully provides a straight-forward alternative that invites researchers to enact upon their sample dependency concerns.

a. Possible applications

One of the several research vectors proposed above was the analysis of social media groups (Admin and invitees). We are very excited for the possibility of researching a naturally structured online community with the tools developed here. We hope that a dataset consisting of records of invitations and personal invitee information would be available in the future. As of now, to the best of our knowledge, there are no public datasets of these kind.

Our model also provides a statistical basis for the covariant analysis of ordinal states in a Markov chain. A medical example for the kind of analysis that could be done

would be “cancer stages”. One can now use our model to measure the association of raising and descending in cancer stages with other binary or ordinal trait.

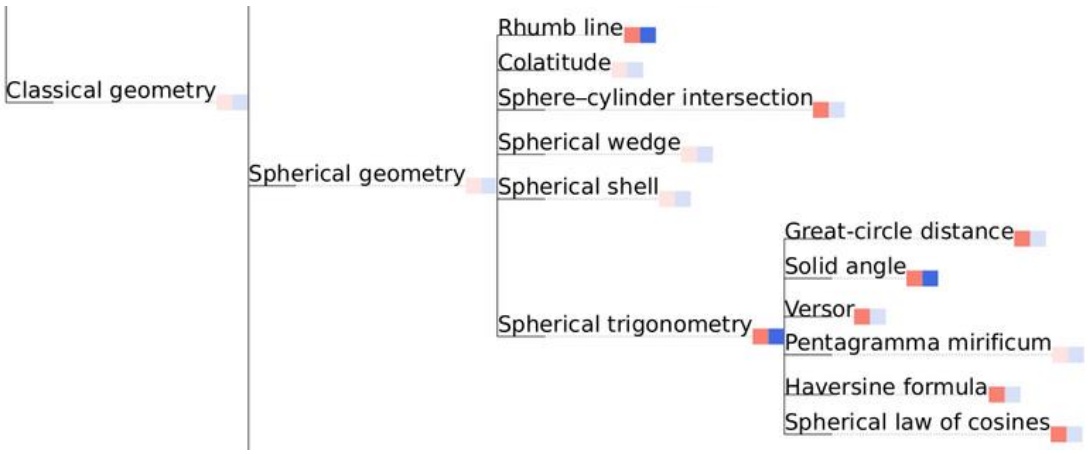
b. Model expansions

A possible extension would be to account for more complex Markov processes. Markov Chain Monte Carlo simulations can be used for the estimation of a general Markov model [43]. However, the computation time needed to use them would need to be properly optimized. Observation Trees can be assigned edge coefficients, which raise/lower the general probability of change proportionally to their length.

We think a logical flow-up work will be to extend our models for continuous Random Walk processes. We have found a probable connection with the CLT Tree-Test and the tests given by Pagel [31], so we have reasonable suspicion for a continuous model to have an analog phenomena to existing continuous trait models such as Phylogenetic Generalized Least Squares [29].

Appendix 1

Subtree from the “Wikipedia:Mathematics” observation tree



0.1 Supplementary 1: “sin” (red) vs “oriented” (blue)

The above figure is a small subtree sampled out of the “Wikipedia:Mathematics” dataset. All nodes represent articles. The red signal is lit the term ‘sin’ is found at the article, and the same goes for blue and ‘oriented’ respectfully.

If a “Covariance score” where to be calculated on this tree, the path:

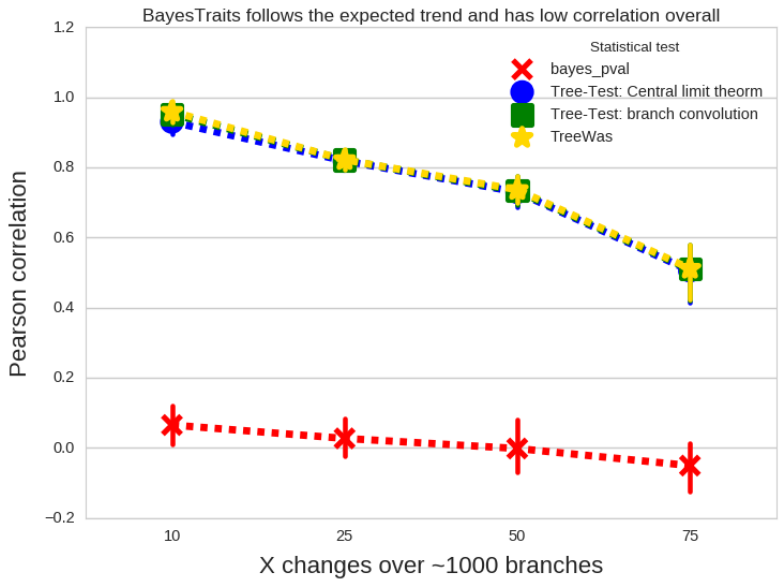
- “Spherical geometry - Spherical trigonometry - Pentagramma mirificum” would contribute +2.
- “Classical geometry - Spherical geometry - Sphere-cylinder intersection” would contribute 0.

Appendix 2

BayesTraits limited comparison

We append a limited comparison between our method and BayesTraits [44]. BaysTraits receives as input a tree and the leaf information of two binary traits. It yields a likelihood estimate of Markov modeling over all possible ancestral reconstructions. After fitting two Markov models, one assuming trait independence and the other not, a likelihood-ratio can be computed. Under the null hypothesis that they are independent, by Wilks' theorem [45], the likelihood-ratio will asymptotically be chi-squared distributed as the sample size approaches infinity. This defines a Likelihood-ratio test, by which significance can be assigned.

The method has theoretical drawbacks. Notably, the asymptotic assumptions are invalid for low transition rates and small populations. The significance cannot be reliably estimated by Wilks' theorem when the ground truth are on the parameter space edge (transition probabilities of 0, or 1). The practical drawbacks involve the estimation of the models. To give a reliable estimation of the best fitting models, substantial computing resources are required. In our tests, using BayesTraits for a single estimation of 100 traits in recommended settings took more than 15 minutes. Nevertheless, we wanted to provide some comparable estimate with it under our simulations in *Chapter II.*



0.1 Supplementary 2: BayesTraits limited simulation

For this, we have limited our simulations: From the M=10k simulated traits, we sampled 100 traits with probability inverse proportional to the ground-truth p-values, providing equal representation. The model fitting for each trait was done in recommended settings. We compare the tests on the p-value of the 100 traits alone, by a non-weighted Pearson correlation with the ground truth. The results are given in *Supplementary 2*.

We see that BayesTraits has the same trend as all other tests, as expected. The higher change rate give room for reconstruction error, which would affect this framework as well. We are nevertheless surprised by the low performance. Overall, as the BayesTraits methods is not promising for large-scale analysis, and not showing significant theoretical or empirical advantages - we end our comparison here.

Bibliography

[1] R. A. Fisher, “On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P,” *J. R. Stat. Soc.*, 1922.

[2] K. Pearson, “Contributions to the Mathematical Theory of Evolution,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 1894.

[3] J. Marchini, L. R. Cardon, M. S. Phillips, and P. Donnelly, “The effects of human population structure on large genetic association studies,” *Nat. Genet.*, 2004.

[4] L. A. Weiss *et al.*, “Genome-wide association study identifies ITGB3 as a QTL for whole blood serotonin,” *Eur. J. Hum. Genet.*, 2004.

[5] D. Welter *et al.*, “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations,” *Nucleic Acids Res.*, 2014.

[6] F. T. Cloak, “Is a cultural ethology possible?,” *Hum. Ecol.*, 1975.

[7] N. Retzlaff and P. F. Stadler, “Phylogenetics beyond biology,” *Theory Biosci.*, vol. 137, no. 2, pp. 133–143, Nov. 2018.

[8] C. J. Howe and H. F. Windram, “Phylomemetics—Evolutionary Analysis beyond the Gene,” *PLoS Biol.*, vol. 9, no. 5, p. e1001069, May 2011.

[9] D. Ramage, A. N. Rafferty, and C. D. Manning, “Random walks for text semantic similarity,” in *ACL-IJCNLP 2009 - TextGraphs 2009: 2009 Workshop on Graph-Based Methods for Natural Language Processing, Proceedings of the Workshop*, 2009.

[10] L. Nakhleh, D. Ringe, and T. Warnow, “Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages,” *Language (Baltim).*, 2005.

[11] A. Adelaar and A. Pawley, *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*. 2009.

[12] J. Nichols and T. Warnow, “Tutorial on Computational Linguistic Phylogeny,” *Lang. Linguist. Compass*, vol. 2, no. 5, pp. 760–820, Sep. 2008.

[13] I. Tëmkin and N. Eldredge, “Phylogenetics and material cultural evolution,” *Curr. Anthropol.*, vol. 48, no. 1, pp. 146–153, Feb. 2007.

[14] W. G. Cochran, “Some Methods for Strengthening the Common χ 2 Tests,” *Biometrics*, 1954.

[15] N. Mantel and W. Haenszel, “Statistical aspects of the analysis of data from retrospective studies of disease,” *J. Natl. Cancer Inst.*, 1959.

[16] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, “Principal components analysis corrects for stratification in genome-wide association studies,” *Nat. Genet.*, 2006.

[17] T. Jombart, S. Devillard, and F. Balloux, “Discriminant analysis of principal components: A new method for the analysis of genetically structured populations,” *BMC Genet.*, 2010.

[18] T. Garland, A. F. Bennett, and E. L. Rezende, “Phylogenetic approaches in comparative physiology,” *Journal of Experimental Biology*. 2005.

[19] W. M. Fitch, “Toward defining the course of evolution: Minimum change for a specific tree topology,” *Syst. Biol.*, vol. 20, no. 4, pp. 406–416, 1971.

[20] B. Chor and T. Tuller, “Maximum likelihood of evolutionary trees: Hardness and approximation,” *Bioinformatics*, 2005.

[21] D. Penny, “Inferring Phylogenies.—Joseph Felsenstein. 2003. Sinauer Associates, Sunderland, Massachusetts.,” *Syst. Biol.*, 2004.

[22] Z. Yang, S. Kumar, and M. Nei, “A new method of inference of ancestral nucleotide and amino acid sequences,” *Genetics*, 1995.

[23] J. P. Huelsenbeck and J. P. Bollback, “Empirical and Hierarchical Bayesian Estimation of Ancestral States,” *Syst. Biol.*, 2001.

[24] F. Lutzoni, M. Pagel, and V. Reeb, “Major fungal lineages are derived from lichen symbiotic ancestors,” *Nature*, 2001.

[25] R. R. Sokal, “A statistical method for evaluating systematic relationships,” *Univ Kans Sci Bull*, 1958.
[26] “The neighbor-joining method: a new method for reconstructing phylogenetic trees.,” *Mol. Biol. Evol.*, 1987.
[27] S. Even and O. Goldreich, “The minimum-length generator sequence problem is NP-hard,” *J. Algorithms*, 1981.
[28] J. Felsenstein, “Phylogenies and the comparative method.,” *Am. Nat.*, 1985.
[29] A. Grafen, “The phylogenetic regression.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 1989.
[30] J. D. Hadfield and S. Nakagawa, “General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters,” *J. Evol. Biol.*, 2010.
[31] D. Barker and M. Pagel, “Predicting functional gene links from phylogenetic-statistical analyses of whole genomes,” *PLoS Comput. Biol.*, 2005.
[32] M. Pagel, “Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters,” *Proc. R. Soc. B Biol. Sci.*, 1994.
[33] W. P. Maddison, “A Method for Testing the Correlated Evolution of Two Binary Characters: Are Gains or Losses Concentrated on Certain Branches of a Phylogenetic Tree?,” *Evolution (N. Y).*, 1990.
[34] P. Martins and T. J. Garland, “PHYLOGENETIC ANALYSES OF THE CORRELATED EVOLUTION OF,” *Evolution*, 1991.
[35] T. Garland, A. W. Dickerman, C. M. Janis, and J. A. Jones, “Phylogenetic analysis of covariance by computer simulation,” *Syst. Biol.*, 1993.
[36] M. R. Farhat, B. J. Shapiro, S. K. Sheppard, C. Colijn, and M. Murray, “A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens,” *Genome Med.*, 2014.
[37] O. Cohen, H. Ashkenazy, D. Burstein, and T. Pupko, “Uncovering the co-evolutionary network among prokaryotic genes,” *Bioinformatics*, 2012.
[38] C. Collins and X. Didelot, “A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination,” *PLOS Comput. Biol.*, vol. 14, no. 2, p. e1005958, Feb. 2018.
[39] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” 1994.
[40] A. Kilgarriff and C. Fellbaum, “WordNet: An Electronic Lexical Database,” *Language (Baltim).*, 2000.
[41] J. P. McCrae, I. Wood, and A. Hicks, “The colloquial WordNet: Extending Princeton WordNet with neologisms,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.
[42] M. J, “Trending YouTube Video Statistics, Version 115.,” *Kaggle*, 2017. [Online]. Available: https://www.kaggle.com/datasnaek/youtube-new.
[43] A. Gupta and J. B. Rawlings, “Comparison of parameter estimation methods in stochastic chemical kinetic models: Examples in systems biology,” *AIChE J.*, 2014.
[44] D. Barker, A. Meade, and M. Page, “Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes,” *Bioinformatics*, 2007.
[45] S. S. Wilks, “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses,” *Ann. Math. Stat.*, 1938.

לאחר מכן אנו מציגים מושג מטרה בדמות מילות "האשטאג" (תג הקבצה) ואת השדה הסמנטי של האשטאגים קשורים. גם במקרה זה המבחן שלנו מפגין את טיבו. בנוסף על כך, אנחנו מציגים כיצד הכלים שלנו יכולים להוסיף על הקיים בתחום.

אחת ממטרות עבודה זו הייתה ליישם אלגוריתמים בעלי מוטיבציה ביולוגית ביישומים לא סטנדרטיים. כחלק ממוטיבציה זאת אנו מציעים לאורך העבודה באופן עקבי פרויקטים אפשריים לקידום המוטיבציה של הקורא. אנו מקווים שהגישה רחבת האופק שלנו תקדם את המודעות לתלות ההיררכית האפשרית במבחנים סטטיסטיים כללים. אנו תקווה כי השיטה שלנו מעניקה חלופה יעילה לקיים בתחום, המזמינה את החוקרים הקוראים לבדוק את האפשרות לתלויות סמויות בתצפיותיהם.

תקציר

בתחום הסטטיסטיקה הקלאסית, כאשר אנו מעריכים את התלות בין שני משתנים בינאריים, נהוג לאסוף מספר רב של תצפיות בלתי-תלויות - ומופעלים מבחנים סטטיסטים מוכרים כדוגמת מבחן "פישר המדויק". אך עבור יישומים רבים בעולם האמתי, הנחת אי-תלות מדגם התצפיות היא לעיתים קרובות שגויה למדי. לדוגמה, בתחום הגנטיקה - ידוע כי כל תצפיות שנאספו אינן עצמאיות, שכן כולן חולקות אב קדמון משותף.

על-מנת לענות על הגורמים המתערבים אשר יכולים לנבוע מהתלות האבולוציונית של מדגמים ביולוגיים - חוקרי הביולוגיה, פיתחו מבחני סטטיסטיים על בסיס עצים. כלים אלה נועדו להתחשב במבנה התלות ההיררכי של המדגם המוכתב על ידי המבנה הטופולוגי של אבות אבותיהם - עץ "משפחה". המבחנים המתקדמים כיום משתמשים בסימולציות מונטה-קרלו על-מנת לקחת בחשבון את מבני התלות הללו. עם זאת, סימולציות אלו דורשות כוח חישובי בלתי מבוטל, אשר מגבילות אותם לשימוש לצורך הערכת קשרים בהיקף מאסיבי.

בעבודתנו אנו מציגים מבחן תלות אנליטי מוכלל, אשר אינו מכיל סימולציה, הנועד עבור מדגמים בעלי מבנה תלות היררכי. המבחן מבוסס על מידול מרקובי של המשתנים, אשר משוערך מתוך התצפיות בעזרת אלגוריתמים ביולוגיים כגון Maximum Parsimony. אנו משתמשים בעומק התיאורטי של המודל המרקובי על מנת לנסח ציון חדש המודד באופן ישיר את השונות המשותפת ואת הכיווניות שלה. המבחן בוחן את השערת האפס לאי-קיום שונות משותפת בשני גרסאות: אחת הכוללת חישוב ישיר של התפלגות הציון הצפויה דרך מספר סופי של צעדי קונבולוציה, ושני אשר משתמש במשפט הגבול המרכזי לשערוך מקורב. בין השאר, המודל בניסוחו המלא מוכלל למשתנים מעל קבוצה סדורה.

לבדיקת טיב המבחן שלנו, אנו מנסחים מספר הנחות ומשווים את ביצועינו תחת סימולציה למסגרת העבודה המייצגת את חזית המחקר כיום. לאחר איסוף התוצאות, אנו מעריכים כי תחת הנחות המודל שלנו - המבחן בגרסת החישוב הישיר מצליח באופן שווה לחזית המחקר. המבחן בגרסת משפט הגבול המרכזי מגיע לביצועים דומים ברוב המקרים, ומציג חולשות ביצועיות באופן המקביל לעבודות קודמות בתחום.

השיטה שלנו ניתנת ליישום באופן בתחומים רבים, שכן מבני תלות היררכיים ניתנים למציאה בסוגי מידע מגוונים כגון: שפה, כלכלה, חברה, רפואה, מדיה וכו'. על-מנת להפגין את החוזק ורמת ההכללה של השיטה שלנו, אנו מציגים ניתוח של כמה מקרי תצפית מתוך אנליזה רחבה של נתונים אשר נאספו מתוך אתרי התוכן "ויקיפדיה", ו"יוטיוב".

שני האתרים הנ"ל מדגימים בהתאמה שני מצבים שונים בהם ניתן להפעיל את המבחן שלנו. באחד, כלל התצפיות שלנו מאורגנות במבנה עץ ומוצבות על פני כלל קדקודיו - כולל הפנימיים. בשני, התצפיות מוצבות בקצוות (עלי) עץ - כאשר אין לנו מידע על הקדקודים הפנימיים. מאגרים אלו נאספו למוטיבציות בלשניות של שערוך יחסי קרבה בין מילים. היקף המילים המוכל במאגרי התוכן הנ"ל הוא בסדר גודל גבוהים, ושערוך הקרבה בין כל זוג דורש חישוב מאסיבי - אך המבחן שלנו מבצע אותו ביעילות רבה.

במסגרת היישום על האתר ויקיפדיה, אנו מציגים שערוך של השדה הסמנטי של מושגים מקצועיים במתמטיקה. הדבר נעשה ע"י איסוף של כאלף מאמרים בתחום תוך זיהוי הקשרים ההיררכיים המייצגים תתי תחומים ומושגים מקומיים בשפה המקצועית. לאחר מכן אנו מציגים מושג מטרה ואת השדה הסמנטי אשר נמצא לו. להשוואה אנחנו מציבים את תוצאותינו לצד אלו של מבחן סטטיסטי אשר מניח אי תלות בין המאמרים. נוסף על כך, אנחנו משתמשים בכלים מתחום עיבוד השפה הטבעית למתן בסיס אחיד להשוואת המבחנים. המבחן שלנו מראה עליונות באופן מובהק סטטיסטית ואיכותנית.

במסגרת היישום על "יוטיוב", אנו מציגים שערוך של השדה הסמנטי סביב שיח אינטרנטי אקטואלי. הדבר נעשה ע"י איסוף של כאלפי סרטוני וידאו ויראליים, ושיערוך הקשר ההיררכי ביניהם ע"י אשכול לפי הז'אנר והיוצר.

המחקר נעשה בהנחייתם המשותפת של פרופסור רועי קישוני מפקולטה למדעי המחשב וכן מהפקולטה לביולוגיה, ופרופסור זוהר יכיני מהפקולטה למדעי המחשב.

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

מבחן-העץ : מבחן סטטיסטי לבדיקת השערת אי-תלות מתצפיות על עץ מכוון

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר מגיסטר למדעים במדעי המחשב

גל נוביץ'

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

אלול תשע"ט חיפה ספטמבר 2019

מבחן-העץ : מבחן סטטיסטי לבדיקת השערת אי-תלות מתצפיות על עץ מכוון

גל נוביץ'

View publication stats