Uncertainty Estimation and Analysis of Categorical Web Data

Davide Ceolin[1], Willem Robert van Hage[2], Wan Fokkink[1], and Guus Schreiber[1]

[1] VU University, Amsterdam, The Netherlands
{d.ceolin,w.j.fokkink,guus.schreiber}@vu.nl
[2] Synerscope B.V., Eindhoven, The Netherlands
willem.van.hage@synerscope.com

Abstract. Web data often manifest high levels of uncertainty. We focus on categorical Web data and we represent these uncertainty levels as first- or second-order uncertainty. By means of concrete examples, we show how to quantify and handle these uncertainties using the Beta-Binomial and the Dirichlet-Multinomial models, as well as how take into account possibly unseen categories in our samples by using the Dirichlet process. We conclude by exemplifying how these higher-order models can be used as a basis for analyzing datasets, once at least part of their uncertainty has been taken into account. We demonstrate how to use the Battacharyya stastistical distance to quantify the similarity between Dirichlet distributions, and use such results to analyze a Web dataset of piracy attacks both visually and automatically.

Keywords: Uncertainty, Bayesian statistics, Non-parametric statistics, Beta-Binomial, Dirichlet-Multinomial, Dirichlet process, Bhattacharyya distance

1 Introduction

The World Wide Web and the Semantic Web offer access to an enormous amount of data and this is one of their major strengths. However, the uncertainty about these data is quite high, due to the multi-authoring nature of the Web itself and to its time variability: some data are accurate, some others are incomplete or inaccurate, and generally, such a reliability level is not explicitly provided.

We focus on the real distribution of these Web data, in particular of categorical Web data, regardless of whether they are provided by documents, RDF [33] statements or other means. Categorical data are among the most important types of Web data, because they include also URIs. We assume that any kind of reasoning that might produce new statements (e.g. subsumption) has already taken place. Hence, unlike for instance Fukuoe et al. [16], that apply probabilistic reasoning in parallel to OWL [32] reasoning, we propose some models to address uncertainty issues on top of that kind of reasoning layers. These models, namely the parametric Beta-Binomial and Dirichlet-Multinomial, and the

non-parametric Dirichlet process, use first- and second-order probabilities and the generation of new classes of observations, to derive safe conclusions on the overall populations of our data, given that we are deriving those from possibly biased samples. These models are chosen exactly because they allow modeling categorical datasets, while taking into account the uncertainty related to the fact that we observe these datasets through samples that are possibly misleading or only partially representative.

Our goal is twofold. On the one hand, we want to show that higher-order probability distributions are useful to model categorical Web datasets while coping with their uncertainty. Hence we compare them with first-order probability distributions and show that taking uncertainty into account is preferable, for instance, when such distributions are used as a basis for prediction. On the other hand, we also show that it is possible to use higher-order probability distributions as basis for data analyses, rather than necessarily focusing on the raw data.

This chapter revises and extends the paper "Estimating Uncertainty of Categorical Web Data" [6], presented at the 7th International Workshop on Uncertainty Reasoning for the Semantic Web at the 10th International Semantic Web Conference 2011. The extension regards mainly the demonstration of use of higher-order probability distributions as a basis for categorical Web data analysis. In particular, we show how to use statistical distances (specifically, the Bhattacharyya statistical distance) to identify patterns and relevant changes in our data.

The chapter continues as follows. First we describe the scope of these models (Section 2), second we introduce the concept of conjugate prior (Section 3), and then two classes of models: parametric and non-parametric (Section 4). We show how it is possible to utilize such models to analyze dataset from the Web (Section 5) and, finally, we discuss the results and conclude (Section 6).

2 Scope of this work

We define here the scope of the work presented in this chapter.

2.1 Empirical evidence from the Web

Uncertainty is often an issue in case of empirical data. This is especially the case with empirical Web data, because the nature of the Web increases the relevance of this problem but also offers means to address it, as we see in this section. The relevance of the problem is related to the utilization of the mass of data that any user can find over the Web: can one safely make use of these data? Lots of data are provided on the Web by entities the reputation of which is not surely known. In addition to that, the fact that we access the Web by crawling, means that we should reduce our uncertainty progressively, as long as we increment our knowledge. Moreover, when handling our sample it is often hard to determine

how representative such a sample is of the entire population, since often we do not own enough sure information about it.

On the other hand, the huge amount of Web data gives also a solution for managing this reliability issue, since it can provide the evidence necessary to limit the risk when using a certain data set.

Of course, even within the Web it can be hard to find multiple sources asserting about a given fact of interest. However, the growing dimension of the Web makes it reasonable to believe in the possibility to find more than one data set about the given focus, at least by means of implicit and indirect evidence.

This work aims to show how it is possible to address the described issues by handling such empirical data, categorical empirical data in particular, by means of the Beta-Binomial, Dirichlet-Multinomial and Dirichlet process models.

2.2 Requirements

Our approach needs to be quite elastic in order to cover several issues, as described below. The non-triviality of the problem comes in a large part from the impossibility to directly handle the sampling process from which we derive our conclusions. The requirements that we need to meet are:

Ability to handle incremental data acquisition The model should be incremental, in order to reflect the process of data acquisition: as long as we collect more data (even by crawling), our knowledge should reflect that increase.

Prudence It should derive prudent conclusions given all the available information. In case not enough information is available, the wide range of possible conclusions derivable should clearly make it harder to set up a decision strategy.

Cope with biased sampling The model should deal with the fact that we are not managing a supervised experiment, that is, we are not randomly sampling from the population. We are using an available data set to derive safe consequences, but these data could, in principle, be incomplete, inaccurate or biased, and we must take this into account.

Ability to handle samples from mixtures of probability distributions The data we have at our disposal may have been drawn from diverse distributions, so we cannot use the central limit theorem, because it relies on the fact that the sequence of variables is identically distributed. This implies the impossibility to make use of estimators that approximate by means of the Normal distribution.

Ability to handle temporal variability of parameters Data distributions can change over time, and this variability has to be properly accounted.

Complementarity with higher-order layers The aim of the approach is to quantify the intrinsic uncertainty in the data provided by the reasoning layer, and, in turn, to provide to higher-order layers (time series analysis, decision strategy, trust, etc.), reliable data and/or metadata.

2.3 Related work

The models adopted here are applied in a variety of fields. For the parametric models, examples of applications are: topic identification and document clustering [12,24], quantum physics [20], and combat modeling in the naval domain [23]. What these heterogeneous fields have in common is the presence of multiple levels of uncertainty (for more details about this, see Section 4.1).

Also non-parametric models are applied in a wide variety of fields. Examples of these applications include document classification [9] and haplotype inference [36]. These heterogeneous fields have in common with the applications mentioned above the presence of several layers of uncertainty, but they also show a lack of prior information about the number of parameters. These concepts are treated in Section 4.2 where the Wilcoxon sign-ranked test [35], used for validation purposes, falls into the non-parametric models class.

Our focus is on the statistical modeling of categorical Web data. The analysis of categorical data is a widespread and well consolidated topic (see, for instance, the work of Davis and Koch [7] or Agresti [1]). About the statistical analysis of Web datasets, Auer et al. [3] present a statement-stream-based approach for gathering comprehensive statistics about RDF datasets that differs from our approach as we do not focus on streams. To our best knowledge, the chosen models have not been applied to categorical Web data yet. We propose to adopt them, because, as the following sections show, they fit the requirements previously listed. Moreover, we see models such as SCOVO [18], RDF Data Cube [8] and VoID [2] as complementary to our work, since these would allow modeling and publishing the results of our analyses.

3 Prelude: conjugate priors

To tackle the requirements described in the previous section, we adopt some Bayesian parametric and non-parametric models in order to be able to answer questions about Web data.

Conjugate priors [17] are the "leit motiv", common to all the models adopted here. The basic idea starts from the Bayes theorem (1): given a prior knowledge and our data, we update the knowledge into a posterior probability.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{1}$$

This theorem describes how it is possible to compute the posterior probability, $P(A|B)$, given the prior probability of our data, $P(A)$, the likelihood of the model, given the data, $P(B|A)$, and the probability of the model itself, $P(B)$.

When dealing with continuous probability distributions, the computation of the posterior distribution by means of Bayes theorem can be problematic, due to the need to possibly compute complicated integrals. Conjugate priors allow us to overcome this issue: when prior and posterior probability distributions belong to the same exponential family, the posterior probability can be obtained by updating the prior parameters with values depending on the observed sample [15].

Exponential families are classes of probability distributions with a density function of the form $f(x) = e^{a(q)b(x)+c(q)+d(x)}$, with q a known parameter and a, b, c, d known functions. Exponential families include many important probability distributions, like the Normal, Binomial, Beta, etc., [11]. So, if X is a random variable that distributes as defined by the function $P(p)$ (for some parameter or vector of parameters p) and, in turn, p distributes as $Q(\alpha)$ for some parameter (or vector of parameters α called "hyperparameter"), and P belongs to the same exponential family as Q,

$$p \sim Q(\alpha),\ X \sim P(p)$$

then, after having observed *obs*,

$$p \sim Q(\alpha')$$

where $\alpha' = f(\alpha, obs)$, for some function f. For example, the Beta distribution is the conjugate of the Binomial distribution. This means that the Beta, shaped by the prior information and by the observations, defines the range within which the parameter p of the Binomial is probably situated, instead of directly assigning to it the most likely value. Other examples of conjugate priors are: Dirichlet, which is conjugate to the Multinomial, and Gaussian, which is conjugate to itself.

Conjugacy guarantees ease of computation, which is a desirable characteristic when dealing with very big data sets as Web data sets often are. Moreover, the model is incremental, and this makes it fit the crawling process with which Web data are obtained, because crawling, in turn, is an incremental process. Both the heterogeneity of the Web and the crawling process itself increase the uncertainty of Web data. The probabilistic determination of the parameters of the distributions adds a smoothing factor that helps to handle this uncertainty.

4 Higher-order probability distributions for modeling categorical Web data

This section presents three higher-order probability distributions that are useful to model uncertain categorical Web data. We present them in order of growing complexity, and then we outline a procedure for employing these models.

4.1 Parametric Bayesian models for categorical Web data

Here we handle situations where the number of categories is known a priori, by using the Dirichlet-Multinomial model and its special case with two categories, i.e. the Beta-Binomial model [15]. Since we handle categorical data, the Binomial and the Multinomial distributions could be the natural choice to model them, depending on whether these data are divided into two or more categories. The Binomial and the Multinomial distributions allow modeling n draws from these datasets. These presume that the frequency of the categories is known, but this is not possible in our case, because we have at our disposal only a data sample

which representativity is unknown. So we still model the data distributions by means of Binomial or Multinomial distributions (depending on the number of categories that we have), but we also model the parameters of these distributions by means of Beta or Dirichlet distributions respectively, since these are conjugated with the Binomial and with the Multinomial distributions. The shape of the Beta and of the Dirichlet distribution is determined by both the size and the distribution of the sample observed. The resulting models (Beta-binomial and Dirichlet-Multinomial) allow us to model the data distribution even though we base our modeling on samples that are uncertain and limited in size.

These models are parametric, since the number and type of parameters is given a priori, and they can also be classified as "empirical Bayesian models". This further classification means that they can be seen as an approximation of a full hierarchical Bayesian model, where the prior hyperparameters are set to their maximum likelihood values according to the analyzed sample.

Case study 1 - Ratio estimation Suppose that a museum has to annotate a particular item I of its collection. Suppose further, that the museum does not have expertise in the house about that particular subject and, for this reason, in order to correctly classify the item, it seeks judgments from outside people, in particular from Web users that provide evidence of owning the desired expertise.

After having collected judgements, the museum faces two possible classifications for the item, $C1$ and $C2$. $C1$ is supported by four experts, while $C2$ by only one expert. We can use these numbers to estimate a probability distribution that resembles the correct distribution of $C1$ and $C2$ among all possible annotations.

A basic decision strategy that could make use of this probability distribution, could accept a certain classification only if its probability is greater or equal to a given threshold (e.g. 0.75). If so, the Binomial distribution representing the sample would be treated as representative of the population, and the sample proportions would be used as parameters of a Bernoulli distribution about the possible classifications for the analyzed item: $P(class(I) = C1) = 4/5 = 0.8$, $P(class(I) = C2) = 1/5 = 0.2$. (A Bernoulli distribution describes the possibility that one of two alternative events happens. One of these events happens with probability p, the other one with probability $1 - p$. A Binomial distribution with parameters n, p represents the outcome of a sequence of n Bernoulli trials having all the same parameter p.)

However, this solution shows a manifest leak. It provides to the decision strategy layer the probabilities for each of the possible outcomes, but these probabilities are based on the current available sample, with the assumption that it correctly represents the complete population of all existing annotations. This assumption is too ambitious. (Flipping a coin twice, obtaining a heads and a tails, does not guarantee that the coin is fair, yet.)

In order to overcome this limitation, we should try to quantify how much we can rely on the computed probability. In other words, if the previously computed probability can be referred to as a "first-order" probability, what we need to compute now is a "second-order" probability [20]. Given that the conjugate prior

for the Binomial distribution representing our data is the Beta distribution, the model becomes:

$$p \sim Beta(\alpha, \beta),\ X \sim Bin(p, n) \tag{2}$$

where $\alpha = \#evidence_{C1} + 1$ and $\beta = \#evidence_{C2} + 1$.

By analyzing the shape of the conjugate prior Beta(5,2), we can be certain enough about the probability of $C1$ being safely above our acceptance threshold. In principle, our sample could be drawn by a population distributed with a $40\% - 60\%$ proportion. If so, given the threshold of acceptance of 0.75, we would not be able to take a decision based on the evidence. However, the quantification of that proportion would only be possible if we know the population. Given that we do not have such information, we need to estimate it, by computing (3), where we can see how the probability of the parameter p being above the threshold is less than 0.5. This manifests the need for more evidence: our sample suggests to accept the most popular value, but the sample itself does not guarantee to be representative enough of the population.

$$P(p \geq 0.75) = 0.4660645,\ p \sim Beta(5, 2) \tag{3}$$

Table 1 shows how the confidence in the value p being above the threshold grows as long as we increase the size of the sample, when the proportion is kept. By applying the previous strategy (0.75 threshold) also to the second-order probability, we still choose $C1$, but only if supported by a sample of size at least equal to 15. Finally, these considerations could also be based on the Beta-Binomial

$\#C1$	$\#C2$	$P(p \geq 0.75)$ $p \sim Beta(\#C1 + 1, \#C2 + 1)$
4	1	0.47
8	2	0.54
12	3	0.88

Table 1: The proportion within the sample is kept, so the most likely value for p is always exactly that ratio. However, given our 0.75 threshold, we are sure enough only if the sample size is 15 or higher.

distribution, which is a probability distribution representing a Binomial which parameter p is randomly drawn from a Beta distribution. The Beta-Binomial summarizes model (2) in one single function (4). We can see from Table 2 that the expected proportion of the probability distribution approaches the ratio of the sample (0.8), as the sample size grows. If so, the sample is regarded as a better representative of the entire population and the Beta-Binomial, as sample size grows, converges to the Binomial representing the sample (see Fig. 1).

$$X \sim BetaBin(n, \alpha, \beta) = p \sim Beta(\alpha, \beta), X \sim Bin(n, p) \tag{4}$$

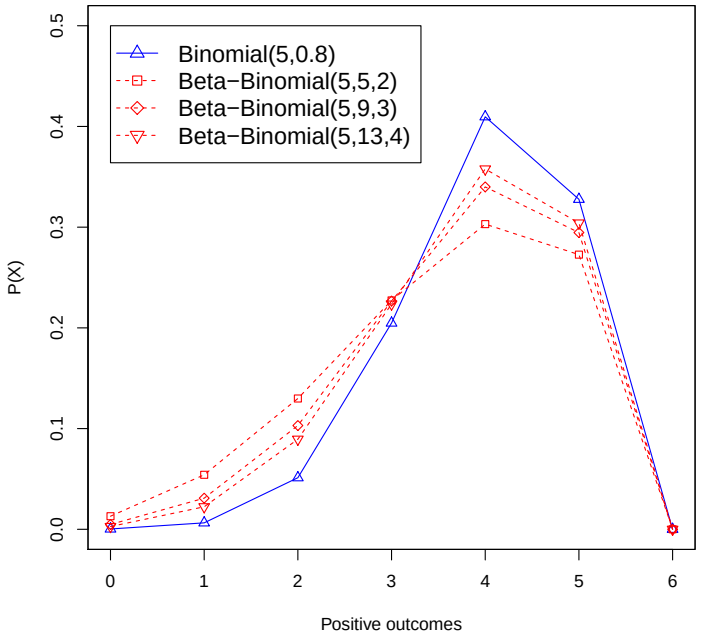


Fig. 1: Comparison between Binomial and Beta-Binomial with increasing sample size. As the sample size grows, Beta-Binomial approaches Binomial.

X	$E(X)$	$p = E(X)/n$
$BetaBin(5,5,2)$	3.57	0.71
$BetaBin(5,9,3)$	3.75	0.75
$BetaBin(5,13,4)$	3.86	0.77

Table 2: The sample proportion is kept, but the "expected proportion" p of Beta-Binomial passes the threshold only with a large enough sample. $E(X)$ is the expected value.

Case study 2 - Confidence intervals estimation The Linked Open Piracy (LOP)[3] is a repository of piracy attacks that happened around the world in the period 2005 - 2011, derived from reports retrieved from the ICC-CCS website.[4] Attack descriptions are provided, in particular covering their type (boarding, hijacking, etc.), place, time, as well as ship type.

Data about attacks is provided in RDF format, and a SPARQL [34] endpoint permits to query the repository. Such a database is very useful, for instance, for insurance companies to properly insure ships. The premium should be related to both ship conditions and their usual route. The Linked Open Piracy repository allows an insurance company to estimate the probability of a ship to be victim of a particular type of attack, given the programmed route. Different attack types imply different risk levels.

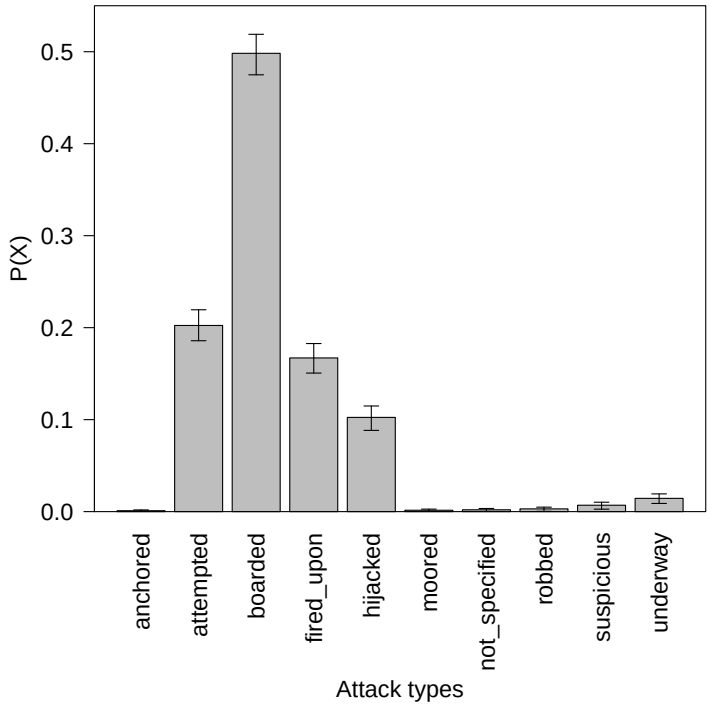


Fig. 2: Attack type proportion and confidence intervals.

However, directly estimating the probability of a new attack given the dataset, would not be correct, because, although derived from data published from an

[3] http://semanticweb.cs.vu.nl/lop
[4] http://www.icc-ccs.org/

official entity like the Chamber of Commerce, the reports are known to be incomplete. This fact clearly affects the computed proportions, especially because it is likely that this incompleteness is not fully random. There are particular reasons why particular attack types or attacks happening in particular zones are not reported. Therefore, beyond the uncertainty about the type of next attack happening (first-order uncertainty), there is an additional uncertainty order due to the uncertainty in the proportions themselves. This can be handled by a parametric model that allows estimating the parameters of a Multinomial distribution. The model that we adopt is the multivariate version of the model described in Subsection 4.1, i.e., the Dirichlet-Multinomial model [12,23,24]:

$$Attacks \sim \text{Multinom}(params),\ params \sim \text{Dirichlet}(\overrightarrow{\alpha}) \tag{5}$$

where $\overrightarrow{\alpha}$ is the vector of observations per attack type (incremented by one unit each, as the α and β parameters of Beta probability distribution). By adopting this model, we are able to properly handle the uncertainty carried by our sample, due to either time variability (over the years, attack type proportions could have changed) or biased samples. Drawing the parameters of our Multinomial distribution from a Dirichlet distribution instead of directly estimating them, allows us to compensate for this fact, by smoothing our attacks distribution. As a result of the application of this model, we can obtain an estimate of confidence intervals for the proportions of the attack types (with 95% of significance level, see Equation (6)). These confidence intervals depend both on the sample distribution and on its dimension (Fig. 2).

$$\forall p \in param, CI_p = (p - \theta_1, p + \theta_2), P(p - \theta_1 \leq p \leq p + \theta_2) = 0.95 \tag{6}$$

4.2 Non-parametric Bayesian models

In some situations, the previously described parametric models do not fit our needs, because they set a priori the number of categories, but this is not always possible. In the previous example, we considered and handled uncertainty due to the possible bias of our sample. The proportions shown by our sample could be barely representative of the entire population because of a non-random bias, and therefore we were prudent in estimating densities, even not discarding entirely those proportions. However, such an approach lacks in considering another type of uncertainty: we could not have seen all the possible categories and we are not allowed to know all of them a priori. Our approach was to look for the prior probability to our data in the n-dimensional simplex, where n is the number of categories, that is, possible attack types. Now such an approach is no more sufficient to address our problem. What we should do is to add yet another hierarchical level and look for the right prior Dirichlet distribution in the space of the probability distributions over probability distributions (or space of simplexes). Non-parametric models differ from parametric models in that the model structure is not specified a priori but is instead determined from data.

The term non-parametric is not meant to imply that such models completely lack parameters, but that the number and nature of the parameters are flexible and not set in advance. Hence, these models are also called "distribution free".

Dirichlet process Dirichlet processes [14] are a generalization of Dirichlet distributions, since they correspond to probability distributions of Dirichlet probability distributions. They are stochastic processes, that is, sequences of random variables (distributed as Dirichlet distributions) which value depends on the previously seen ones. Using the so-called "Chinese Restaurant Process" representation [26], it can be described as follows:

$$X_n = \begin{cases} X_k^* & \text{with probability } \frac{num_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } H & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases} \tag{7}$$

where H is the continuous probability measure ("base distribution") from which new values are drawn, representing our prior best guess. Each draw from H returns a different value with probability 1. α is an aggregation parameter, inverse to the variance: the higher α, the smaller the variance, which can be interpreted as the confidence value in the base distribution H: the higher the α value is, the more the Dirichlet process resembles H. The lower the α is, the more the value of the Dirichlet process tends to the value of the empirical distribution observed. Each realization of the process is discrete and is equivalent to a draw from a Dirichlet distribution, because, if

$$G \sim DP(\alpha, H) \tag{8}$$

is a Dirichlet process, and $\{B\}_{i=1}^n$ are partitions of the domain of H, S, we have that

$$(G(B_1)...G(B_n)) \sim Dirichlet(\alpha H(B_1)...\alpha H(B_n)) \tag{9}$$

If our prior Dirichlet process is (8), given (9) and the conjugacy between Dirichlet and Multinomial distribution, our posterior Dirichlet process (after having observed n values θ_i) can assume one of the following representations:

$$(G(B_1)...G(B_n))|\theta_1...\theta_n \sim Dirichlet(\alpha H(B_1) + n_{\theta_1}...\alpha H(B_n) + n_{\theta_n}). \tag{10}$$

$$G \mid \theta_1...\theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}\frac{\Sigma_{i=1}^n \delta_{\theta_i}}{n}\right) \tag{11}$$

where δ_{θ_i} is the Dirac delta function [10], i.e., the function having density only in θ_i. The new base function is therefore a merge of the prior H and the empirical distribution, represented by means of a sum of Dirac delta's. The initial status of a Dirichlet process posterior to n observations, is equivalent to the nth status of the initial Dirichlet process that produced those observations (using the De Finetti theorem [19]).

The Dirichlet process, starting from a (possibly non-informative) "best guess", as long as we collect more data, approximates the real probability distribution.

Hence, it correctly represents the population in a prudent (smoothed) way, exploiting conjugacy like the Dirichlet-Multinomial model, that approximates well the real Multinomial distribution only with a large enough data set (see Subsection 4.1). The improvement of the posterior base distribution is testified by the increase of the α parameter, proportional to the number of observations.

Case study 3: Unseen categories generation We aim at predicting the type distributions of incoming attack events. In order to build an "infinite category" model, we need to allow for event types to be randomly drawn from an infinite domain. Hence, we map already observed attack types with random numbers in [0..1] and, since all events are a priori equally likely, then new events are drawn from the Uniform distribution, $U(0,1)$, that is our base distribution (and is a measure over [0..1]). The model is:

- $type_1 \sim DP(\alpha, U(0,1))$: the prior over the first attack type in region R;
- $attack_1 \sim Categorical(type_1)$: type of the first attack in R during $year_y$.

After having observed $attack_{1...n}$ during $year_y$, our posterior process becomes:

$$type_{n+1} \mid attack_{1...n} \sim DP\left(\alpha + n, \frac{\alpha}{\alpha+n}U(0,1) + \frac{n}{\alpha+n}\frac{\Sigma_{i=1}^{n}\delta_{attack_i}}{n}\right)$$

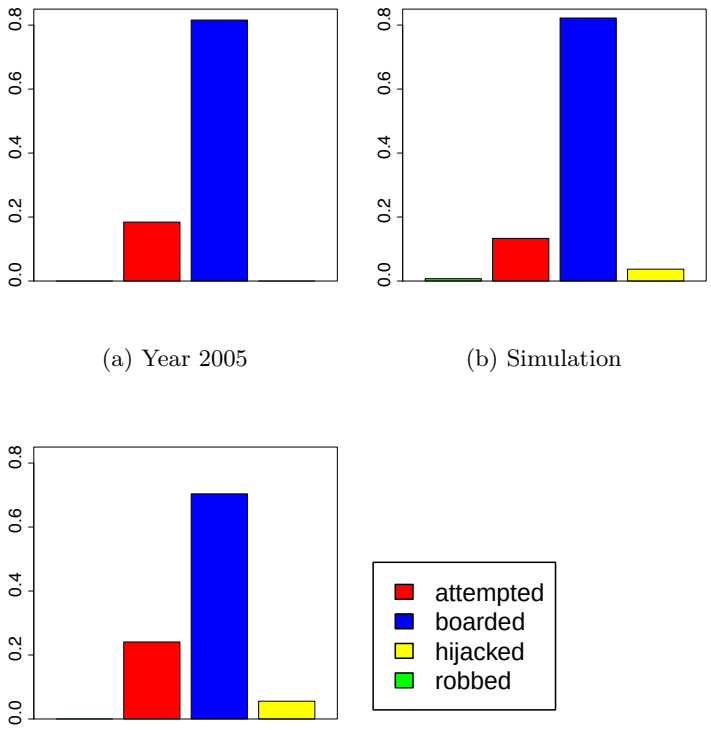
where α is a low value, given the low confidence in $U(0,1)$, and $type_{n+1}$ is the prior of $attack_{n+1}$, that happens during $year_{y+1}$. A Categorical distribution is a Bernoulli distribution with more than two possible outcomes (see Subsection 4.1).

Results Focusing on each region at time, we simulate all the attacks that happened there in $year_{y+1}$. Names of new types generated by simulation are matched to the actual $year_{y+1}$ names, that do not occur in $year_y$, in order of decreasing probability. The simulation is compared with a projection of the proportions of $year_n$ over the actual categories of $year_{n+1}$. The comparison is made by measuring the distance of our simulation and of the projection from the real attack types proportions of $year_{y+1}$ using the Manhattan distance [22]. This metric simply sums, for each attack type, the difference between the real $year_{y+1}$ probability and the one we forecast. Hence, it can be regarded as an error measure. Table 3 summarizes the results over the entire dataset.[5] Our simulation reduces the distance (i.e. the error) with respect to the projection, as confirmed by a Wilcoxon signed-rank test [35] at 95% significance level. (This non-parametric statistical hypothesis test is used to determine whether one of the means of the population of two samples is smaller/greater than the other.) The simulation improves when a large amount of data is available and the category cardinality varies, as in case of Region India, which results are reported in Fig. 3 and 4a.

[5] The code is available at `http://trustingwebdata.org/books/URSW_III/DP.zip`.

	Simulation	Projection
Average distance	**0.29**	0.35
Variance	**0.09**	0.21

Table 3: Averages and variances of the prediction errors. The simulation gets a better performance.

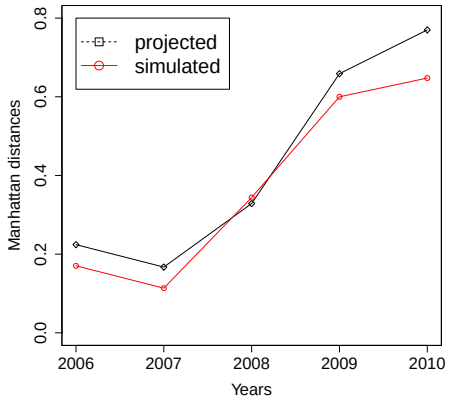


(a) Year 2005

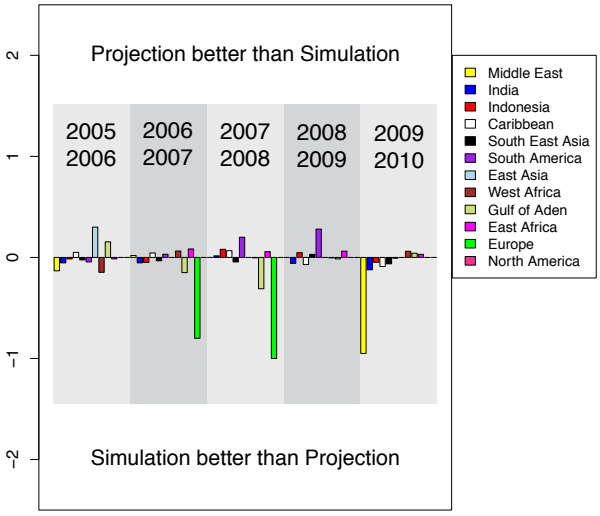
(b) Simulation

(c) Year 2006

Fig. 3: Comparison between the projection forecast and the simulation forecast with the real-life year 2006 data of region India.



(a) Errors



(b) Distances differences

Fig. 4: Error distance from real distribution of the region India (Fig. 4a) and differences of the error of forecast based on simulation and on projection (Fig. 4b). Positive difference means that the projection predicts better than our simulation.

4.3 Model selection and utilization

Here we provide generic indication about the choice and use of the models described before.

Model selection The models presented above are closely related each other, since each of them represents a generalization of the preceding model. Algorithm 1 is the algorithm that we propose for choosing the right model to apply when handling categorical Web data. It is rather simple and, under the assumption that we are handling categorical data, determines the choice of the model to use based on the number of categories that are known to be present in the data.

```
if the number of categories is known then
    if the number of categories is two then
        return Beta-Binomial
    else
        return Dirichlet-Multinomial
    end
else
    return Dirichlet Process
end
```

Algorithm 1: Model Selection Algorithm.

Model building Once the model has been selected, we build it based on the observations at our disposal as follows:

Beta-Binomial The Beta-binomial model has three parameters: n, i.e. the number of draws performed and α and β, that are the frequencies of the two categories. In case no prior knowledge is available, we add the uninformative prior 1 to each frequency parameter. Otherwise, we add the prior frequency to each parameter.

Dirichlet-Multinomial This model has a vector $\overrightarrow{\alpha}$ of frequency parameters, plus the same n parameter indicating the number of draws to perform, as above. The frequency parameters need to be populated with the absolute frequencies observed. In case no prior knowledge is available, we add the uninformative prior 1 to each frequency parameter. Otherwise, we add the prior frequency to each parameter.

Dirichlet Process The Dirichlet Process is determined by two parameters: the concentration parameter α and the base distribution H. If no prior information is available, we set $\alpha = 1$ and $H = U(0, 1)$, where U stands for the Uniform probability distribution. Then, after n categorical observations, we obtain the process described in Equation 11.

Model utilization In the examples above, we used the process for prediction. In the following analyses we use them for comparison. To compare models, we compute similarity measures between probability distributions and analyze them. We present a detailed description of this utilization of the models in the following section. To perform predictions, i.e., to draw from the probability distributions, we proceed as follows:

Beta-binomial (and Dirichlet-multinomial) Randomly draw a parameter p (or a vector of parameters $\overrightarrow{p}$ in the case of the Dirichlet-multinomial) from a Beta (Dirichlet) distribution shaped by the frequency parameters α and β ($\overrightarrow{\alpha}$) described above. Then, randomly draw from a Binomial (Dirichlet) distribution shaped by the parameter p ($\overrightarrow{p}$).

Dirichlet Process Draw from the Dirichlet distribution described above and in Equation 11. If the drawn value has not been observed yet, then draw again from the base distribution H. Then update the process in order to obtain a new Dirichlet distribution representing its updated state.

5 Analyzing datasets using higher-order probability distributions

In the previous sections we have shown that higher-order probability distributions are useful to model Web data and account for their uncertainty. Here we want to show that higher-order probability distributions, despite the fact that they introduce a computational layer in the data management process, are easily utilizable as a basis for data analyses. The analyses presented here aim at showcasing how a data analyst could use the models presented before to derive insights from the data, for summarizing them and for extracting potentially useful information from large datasets. Besides the uncertainty management advantage, these models provide a means to abstract the data we analyze, thus allowing us to identify interesting patterns and regularities that would be hidden otherwise.

We apply our analyses on the LOP dataset introduced before. In the previous section, and in particular in Case Study 3 (Subsection 4.2), we represent piracy attacks spread over the world by means of Dirichlet processes that "generate" the attacks over time. Each step of the Dirichlet process is represented by a Dirichlet distribution. We analyze the type distribution of the attacks with respect to time and regions. So, we use the data at our disposal to build one Dirichlet distribution per region per year, to represent the attack types distributions while taking into account the uncertainty in the data. Then, we use a statistical similarity to measure the likeness of distributions over time and regions. Clearly, we could have used different methods (e.g., mixture models), but we prefer this approach for its flexibility and simplicity.

5.1 Bhattacharyya distance

We adopt the Bhattacharyya distance [4] to quantify the similarity between attack types distributions. The Bhattacharyya distance is a measure of divergence

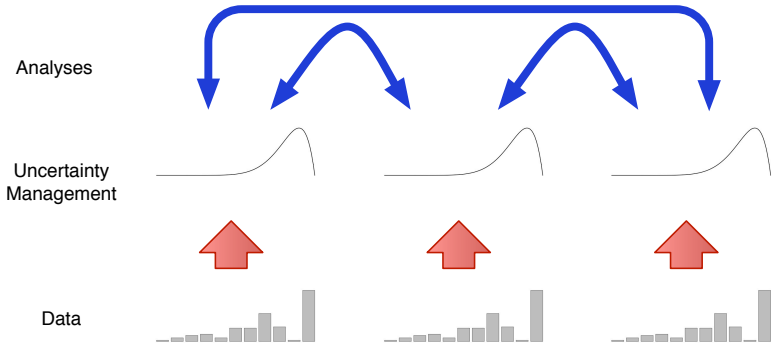


Fig. 5: Data abstraction and analysis overview.

between probability distributions, that allows measuring the dissimilarity between two continuous or discrete probability distributions. As such, it goes from zero (when the compared distributions are identical) to infinite (when there is no overlap between the compared distributions). For continuous probability distributions, it is defined as follows:

$$D_B(dist_a, dist_b) = -\ln\left(\int \sqrt{dist_a(x)dist_b(x)}dx\right)$$

When applied to the Dirichlet Distributions (Rauber et al. [28]), the Bhattacharyya distance becomes:

$$D_B(Dir_a(x_1,\ldots,x_n), Dir_b(y_1,\ldots,y_n)) =$$
$$\Gamma\left(\frac{1}{2}\sum_{i\in\{1,\ldots,n\}} x_i + \frac{1}{2}\sum_{i\in\{1,\ldots n\}} y_i\right) + \frac{1}{2}\sum_{i\in\{1,\ldots,n\}}(\Gamma(x_i)) + \frac{1}{2}\sum_{i\in\{1,\ldots,n\}}(\Gamma(y_i)) -$$
$$\sum_{i\in\{1\ldots n\}}\left(\Gamma\left(\frac{1}{2}(x_i+y_i)\right)\right) - \frac{1}{2}\Gamma\left(\sum_{i\in\{1,\ldots,n\}}(x_i)\right) + \frac{1}{2}\Gamma\left(\sum_{i\in\{1,\ldots,n\}}(y_i)\right)$$

An advantage of the adopted approach is that the computation of the Bhattacharyya distance is particularly convenient. The only change we apply to the distance is to apply the logarithm base 2 to the result of the measure when the value is different from zero. This allows us to handle large numbers without any problem. Thus, the formula becomes:

$$sim(Dir_a, Dir_b) = \begin{cases} 0, & \text{if } D_B(Dir_a, Dir_b) = 0 \\ log_2(D_B(Dir_a, Dir_b)), & \text{otherwise} \end{cases}$$

5.2 Analysis of the distribution of piracy attacks

We measure the Bhattacharyya distance for all the possible combinations of regions and years in the LOP dataset at our disposal. Since the set obtained by computing such similarities is rather big, we split it in two manners: first we look at the similarity of the attack type distributions of different regions, year by year, and second, we analyze the temporal evolution of such similarities, region by region. In this manner, we aim at identifying: (1) similarities in the type distribution across different regions and, (2) patterns related to the temporal distribution of attack types across different regions.

Attack type distribution analysis per year We start by grouping our distances per year, and by analyzing their distribution across different regions of the world. In this manner, we aim at identifying similarities between regional attack distributions, while taking into account the temporal evolution of the attacks. Fig. 6 shows six heatmaps representing the similarity between all possible combinations of regions for the years considered. Here, we can identify a few peculiar facts. For instance, Indonesia happens to be a region particularly different from the others (due to the presence of a high number of "boarded" and "attempted" attacks in that region), although this difference reduces in the last years of the period considered. With respect to previous analyses based on the actual piracy attacks counts [31], this difference is higher. From a manual investigation, we note that, besides a difference in the data distribution, Indonesia presents a difference in the total number of attacks registered. The higher-order models that we propose allow taking into account both aspects at the same time. Also, we note that the region that comprises India and Bengal differentiates from the rest in the first three years considered, while an important change in the similarity trend happens in Gulf of Aden in 2008 and continues afterwards.

Attack type distribution analysis per region Fig. 7 and 8 show a series of heatmaps representing the yearly distribution of piracy attacks, grouped by region. Here we can see, for instance, that North America (and, in part, also Europe) is characterized by being quite uniform in its distributions (thanks also to the fact that piracy attacks are quite rare in this region). Also, 2009 and 2010 are two years representing a changing point in several regions (e.g., Gulf of Aden, South America). Given the extension of such changes, we suppose this might be due to one or more events causing the global distribution of piracy attacks to change, although we are not aware of any.

5.3 Automating piracy attacks analysis

The previous section proposes a combination of automatic and visual analysis of the data. Here we finally propose a procedure for automating the process of identifying potential interesting pieces of data in our datasets.

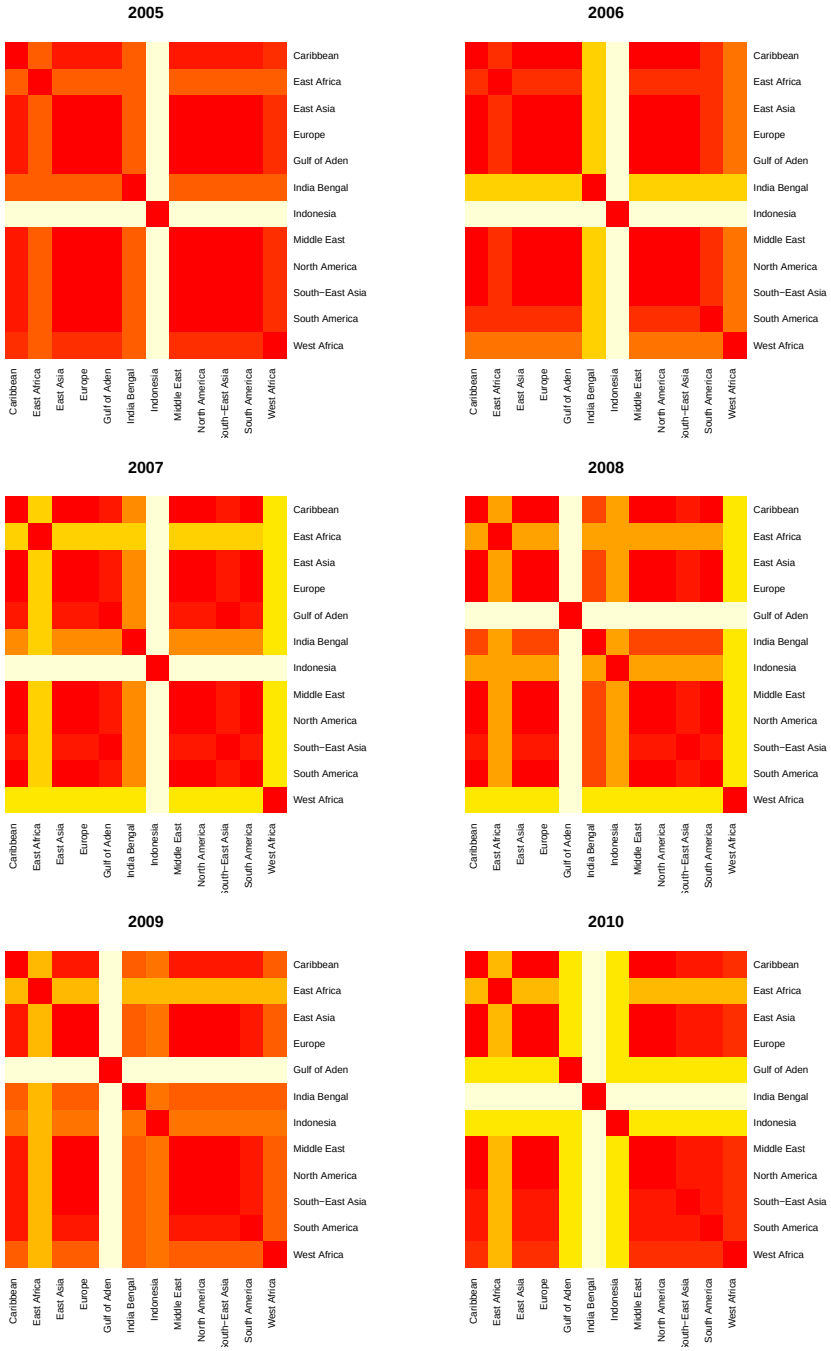


Fig. 6: Heatmaps of the similarity of attack type distributions of different regions of the world, divided by years.

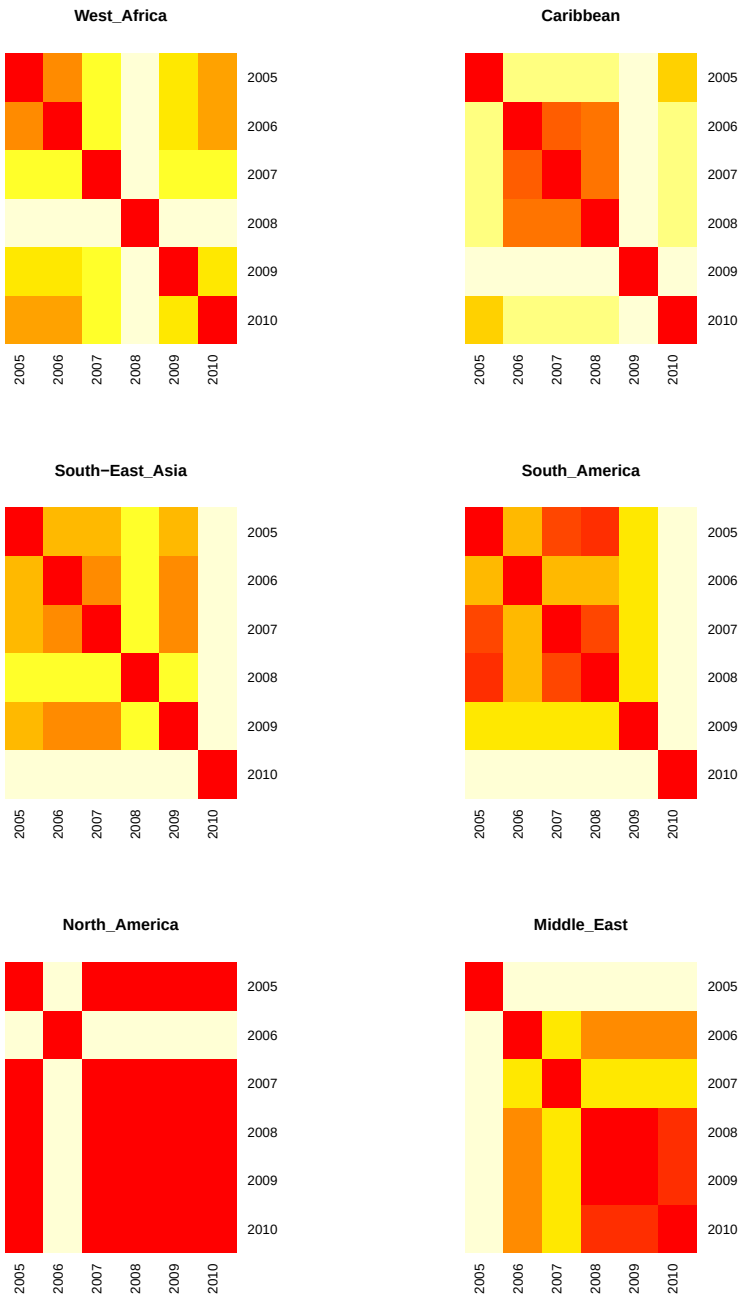


Fig. 7: Attack type distributions of different regions of the world.

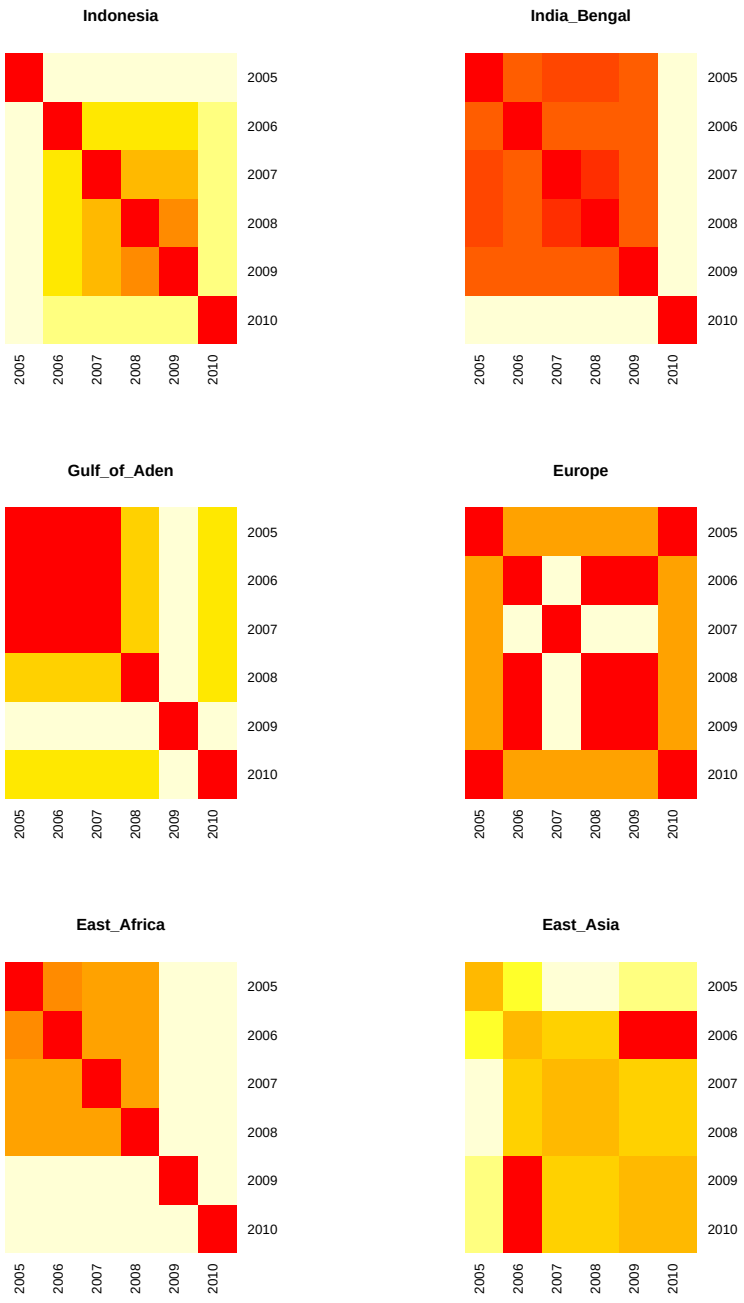


Fig. 8: Attack type distributions of different regions of the world.

Data: A dataset (*dataset*) of piracy attacks
Result: A set of change points in the piracy attacks dataset
Data_Analysis *dataset*
  dm ← `compute_distance_matrix` (*dataset*);
  agg_data ← `aggregate_data` (*dm*);
  res ← `changepoint` (agg_data);
  return *res*;

Algorithm 2: Data Analysis Algorithm.

compute_distance_matrix This procedure computes a similarity matrix that contains a distance between higher-order probabilistic models (e.g., the Bhattacharyya distance defined above in our case).

aggregate_data This procedure aggregates the data with respect to a feature of interest. For instance, we can aggregate the data by time or region of the world, to see variation of attack types over either time or space.

changepoint This procedure relies on the R package "ChangePoint" [21] and identifies points in the aggregated series of piracy attacks distributions that significantly differ from the rest. In particular, we make use of the *cpt.meanvar* function of the package, that determines the change point on the basis of a change in the mean and variance with respect to the rest of the series.

Results We run the above procedure on the LOP dataset, and we obtained:

Regional Aggregation India and Bengal (in 2005), Indonesia (in 2006), West Africa (in 2007, 2008 and 2010) are the regions identified as change points;
Yearly Aggregation 2005 (in East Africa, North America and West Africa), 2006 (in Europe and Gulf of Aden), 2007 (in East Asia), 2008 (in Caribbean, India and Bengala, Indonesia, Middle East, South-East Asia and South America) are the years indicated as change point.

The results of the visual and of the automated analyses present an overlap. The differences are possibly due to the change point detection algorithm chosen. The use of other algorithms will be investigated in the future.

6 Conclusions and future work

We propose a series of higher-order probabilistic models to manage Web data and we show that these models allow us to take into account the inner uncertainty of these data, while providing a probabilistic model that allows reasoning about the data. We demonstrate that these models are useful to handle the uncertainty present in categorical Web data by showing that predictions based on them are more accurate than predictions based on first-order models. Higher-order models allow us to compensate the fact that they are based on limited or possibly

biased samples. Moreover, we show how these models can be adopted as a basis for analyzing the datasets that they model. In particular, we show through a case study, how to exploit statistical distances of probability distributions to analyze the data distribution to identify interesting points within the dataset. This kind of analysis can be used by data analysts to have an insight about the dataset, possibly to be combined with domain knowledge. We propose two kinds of analyses, one based on visual interpretation of heatmaps, the other one based on automatic determination of change points by means of a procedure that we introduce. The results obtained with these two analyses are partially overlapping. Differences are possibly due to the choice of the change point detection algorithm.

In the future, we aim at expanding this work in two directions. Firstly, we plan to extend the set of models adopted, to deal with concrete domain data (e.g. time intervals, by means of the Poisson process [15]), and more sophisticated to improve the uncertainty management part (e.g., Mixture Models [27], Nested [29] and Hierarchical Dirichlet processes [30] and Markov Chain Monte Carlo algorithms [13,25]). Secondly, we will work on the generalization of the data analysis procedures, by combining this work with a previous work on the analysis of the reliability of open data [5] and extending both of them.

Acknowledgements This research was partially supported by the Data2Semantics Media project in the Dutch national program COMMIT.

References

1. Alan Agresti. *Categorical Data Analysis*. Wiley, 3rd edition, 2013.
2. Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing linked datasets with the void vocabulary. Technical report, W3C, 2011.
3. Sören Auer, Jan Demter, Michael Martin, and Jens Lehmann. Lodstats – an extensible framework for high-performance dataset analytics. In *EKAW*, pages 353–362. Springer, 2012.
4. A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.
5. D. Ceolin, L. Moreau, K. O'Hara, W.R. van Hage, W.J. Fokkink, V. Maccatrozzo, G. Schreiber, and N. Shadbolt. Two procedures for estimating the reliability of open government data. In *IPMU*, pages 15–24. Springer, 2014.
6. D. Ceolin, W.R. van Hage, W.J. Fokkink, and G. Schreiber. Estimating Uncertainty of Categorical Web Data. In *URSW*, pages 15–26. CEUR-WS.org, November 2011.
7. Gary Koch Charles Davis. *Categorical Data Analysis Using SAS*. SAS Institute, 3rd edition, 2012.
8. Richard Cyganiak, Dave Reynolds, and Jeni Tennison. The rdf data cube vocabulary. Technical report, W3C, 2014.
9. M. Davy and J. Tourneret. Generative supervised classification using dirichlet process priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1781–1794, 2010.
10. P. Dirac. *Principles of quantum mechanics*. Oxford at the Clarendon Press, 1958.

11. Andersen E. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65:1248–1255, 9 1970.
12. C. Elkan. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *ICML*, pages 289–296. ACM, 2006.
13. M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1994.
14. T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.
15. D. Fink. A Compendium of Conjugate Priors. Technical report, Cornell University, 1995.
16. A. Fokoue, M. Srivatsa, and R. Young. Assessing trust in uncertain information. In *ISWC*, pages 209–224. Springer, 2010.
17. R. Schlaifer H. Raiffa. *Applied statistical decision theory*. M.I.T. Press, 1968.
18. Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayers. SCOVO: Using Statistics on the Web of Data. In *ESWC*, pages 708–722. Springer, 2009.
19. M. Hazewinkel. *Encyclopaedia of Mathematics*, chapter De Finetti theorem. Springer, 2001.
20. J. Hilgevoord and J. Uffink. Uncertainty in prediction and in inference. *Foundations of Physics*, 21:323–341, 1991.
21. Rebecca Killick and Idris A. Eckley. *changepoint: An R Package for Changepoint Analysis*, 2013. `http://cran.r-project.org/package=changepoint`.
22. E. F. Krause. *Taxicab Geometry*. Dover, 1987.
23. P. Kvam and D. Day. The multivariate polya distribution in combat modeling. *Naval Research Logistics (NRL)*, 48(1):1–17, 2001.
24. R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In *ICML*, pages 545–552. ACM, 2005.
25. R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
26. J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102(2):145–158, 1995.
27. Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.
28. T.W. Rauber, A. Conci, T. Braun, and K. Berns. Bhattacharyya probabilistic distance of the dirichlet density and its application to split-and-merge image segmentation. In *WSSIP08*, pages 145–148, 2008.
29. A. Rodriguez, D. B. Dunson, and A. E. Gelfand. The nested dirichlet process. *Journal of the American Statistical Assoc.*, 103(483):1131–1144, September 2008.
30. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Assoc.*, 101(476):1566–1581, 2006.
31. W. R. van Hage, M. van Erp, and V. Malaisé. Linked open piracy: A story about e-science, linked data, and statistics. *J. Data Semantics*, 1(3):187–201, 2012.
32. W3C. OWL Reference, August 2011. http://www.w3.org/TR/owl-ref/.
33. W3C. Resource Definition Framework, August 2011. http://www.w3.org/RDF/.
34. W3C. SPARQL, August 2011. http://www.w3.org/TR/rdf-sparql-query/.
35. F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.
36. E. Xing. Bayesian Haplotype Inference via the Dirichlet Process. In *ICML*, pages 879–886. ACM Press, 2004.