Biometric Zoos: Theory and Experimental Evidence*

Mohammad Nayeem Teli[1†] J. Ross Beveridge[1] P. Jonathon Phillips[2] Geof H. Givens[3]
David S. Bolme[1] Bruce A. Draper[1]

[1] Computer Science, Colorado State University, Fort Collins, CO, USA
[2] Information Access Division, NIST, Gaithersburg, MD, USA
[3] Statistics, Colorado State University, Fort Collins, CO, USA

Abstract

Several studies have shown the existence of biometric zoos. The premise is that in biometric systems people fall into distinct categories, labeled with animal names, indicating recognition difficulty. Different combinations of excessive false accepts or rejects correspond to labels such as: Goat, Lamb, Wolf, etc. Previous work on biometric zoos has investigated the existence of zoos for the results of an algorithm on a data set. This work investigates biometric zoos generalization across algorithms and data sets. For example, if a subject is a Goat for algorithm A on data set X, is that subject also a Goat for algorithm B on data set Y? This paper introduces a theoretical framework for generalizing biometric zoos. Based on our framework, we develop an experimental methodology for determining if biometric zoos generalize across algorithms and data sets, and we conduct a series of experiments to investigate the existence of zoos on two algorithms in FRVT 2006.

1. Introduction

This paper revisits some fundamental questions raised originally by Doddington et al. [2] concerning the role of the individual in frustrating the system engineer charged with the development of reliable biometric authentication. As we will review in more detail shortly, Doddington put forward a captivating hypothesis that different kinds of people were responsible for complicating the task of biometric authentication. In particular, that people could be categorized as Sheep, Goats, Lambs and Wolves. Numerous works have subsequently pushed on this idea from several directions, refining the underlying definitions and arriving at different conclusions with respect to the relative importance of the individual person. Ross et al. [8] used the Doddington zoo effect to categorize the users into multiple categories. Subsequently the weak users are required to provide additional information to improve the overall performance of the system. Poh et al. [6] looked at the entire user population and normalized the score of each user to a standard form. They presented a framework to determine whether the fusion of the output of one or more systems is better than any one output, per user. In another work, Poh and Kitler [7] concluded that biometric zoos are algorithm dependent.

Yager and Dunstone [9] in particular took on the question of how much of what makes biometric authentication difficult may be attributed to there being people who are hard to recognize. In their paper they concluded that the differences observed in recognition difficulty between individual people on a particular data set were attributable to factors other than the identity of the people. They concluded the notion that people fit neatly into distinct categories of difficulty is essentially false.

We believe there is more going on with regard to this fundamental question than either prior works suggest, and we take up the matter in the context of human face recognition. Looking at the performance of two top commercial algorithms in the Face Recognition Vendor Test (FRVT) 2006, we first use Doddington's original definitions to label people on different data sets, and then use statistical hypothesis testing to assess whether these labels are capturing any intrinsic characteristic of the people themselves. We then repeat the analysis for Yager and Dunstone's categories.

Our findings are summarized as follows. Any simplistic notion that variations in recognition performance are caused primarily by intrinsic differences between people is wrong. In this respect, our findings in this paper are consistent with

*The work was funded in part by the Technical Support Working Group (TSWG) under Task SC-AS-3181C. Jonathon Phillips was supported by the Department of Homeland Security, Director of National Intelligence, Federal Bureau of Investigation and National Institute of Justice. The identification of any commercial product or trade name does not imply endorsement or recommendation by Colorado State University or the National Institute of Standards and Technology.

†nayeem@cs.colostate.edu

978-1-4577-1359-0/11/$26.00 ©2011 IEEE

prior findings of others such as Yager and Dunstone [9] . There are factors involving lighting, setting, expression, etc. that more directly influence recognition difficulty [1].

However, within the context of a single algorithm, it is equally wrong to presume that personal identity has nothing to do with recognition difficulty. We demonstrate this second finding using a rigorous statement of the underlying hypothesis that personal identity does not matter, and this formulation and the associated results represent the novel contribution of our work.

This statistical hypothesis testing approach is used as the basis for a new framework for describing and testing for the existence of different levels of biometric zoo. At the bottom is a zeroth-order zoo implying only that people may be labeled as animals in a single experiment. A first-order zoo exists when personal identity is shown to matter for other data drawn from the same scenario, and higher-order zoos pertain to greater levels of generalization.

The remainder of our paper is organized as follows. Section 2 introduces our hierarchy of zoos and outlines the evidence required for the existence of a zoo. Section 3 reviews the original Doddington's zoo and Section 4 illustrates our framework applied to that zoo in the context of frontal face recognition. Section 5 reviews the zoo proposed by Yager and Dunstone and Section 6 applies our framework to the animals in the that zoo. Finally, we summarize our results in the last section.

2. A Framework for a Hierarchy of Zoos

We introduce a hierarchy of zoos generalized with respect to algorithms and data. A zeroth-order zoo focuses solely on the distribution of match and non-match scores for a particular test of an algorithm. In other words, the scores for a particular algorithm on a particular data set. For example, Goats are people with unusually low match scores, and for any algorithm and data set one can sort people by their average match score and find the lowest 2 or 5%.

A first-order zoo is one that generalizes to new data drawn from a common class of data, a scenario. In the framework we are proposing, the existence of a first-order zoo is tested through a series of repeated trials in which a larger data set is randomly sampled to produce two partitions. These partitions have no images in common, but are otherwise representative of the larger set from which they are drawn. If the zoo labels given to people have any meaning in terms intrinsic difficulty associated with the individual person, then the same people should show up with the same label in both partitions.

Our framework defines a first-order zoo more precise by equating it with a statistical hypothesis test. If the existence of a first-order zoo means that the labels given to people are associated with intrinsic properties of those people that generalize given new images, then the number of people sharing a common label for two random draws of data should exceed that predicted by chance. Therefore, the null-hypothesis is that the number of people sharing a given label is governed solely by chance. Under the null-hypothesis, if k out of n people are labeled as Goats in each of two partitions of the data, the expected number labeled Goats in both partitions is governed by a hypergeometric distribution.

To test for the existence of a first-order zoo, we can randomly partition our larger data set and each time count the number people given the same label in each partition. Doing this for $t(= 50)$ randomized trials, we get a count of how often zero people are assigned the same label, one person is assigned the same label, and so on. Then, a χ^2 test is used to determine the probability the observed number of people given the same zoo label arises from the hypergeometric distribution associated with our null-hypothesis. This probability is the p-value associated with the null-hypothesis that zoo labels have no relation to personal identity. Existence of a first-order zoo is now a matter of accepting or rejecting the null hypothesis based upon a p-value.

This same hypothesis testing framework extends to higher order zoos. In particular, we define a second-order zoo to exist when the zoo labels assigned to people for a given class of data generalize across algorithms. Likewise, a second order zoo exists when the labels assigned to people for a particular algorithm generalize across two different classes of data. In our experiments below, we will specifically consider two classes of face recognition image data, controlled to uncontrolled lighting and uncontrolled to uncontrolled lighting.

The hierarchy may be extended. A third-order zoo implies generalization across both algorithms and data sets. A fourth-order zoo implies generalization across biometric modalities. All of these are of interest in the long term, but here we consider only first-, second- and third-order zoos in the context of still image face recognition.

3. Doddington's Zoo

Doddington et al. [2] introduced the concept of the biometric zoo and based the definition of animals on either extreme match or extreme non-match scores.

3.1. Sheep, Goats, Lambs and Wolves

Doddington et al. [2] defined Sheep, Goats, Lambs and Wolves in the context of speaker recognition systems, and their definitions may be summarized as follows:

- Sheep: A person who is a Sheep produces a biometric that matches well to other biometrics of themselves and poorly to those of other people. As such, Sheep generate fewer false accepts and rejects than average.
- Goats: A person who is a Goat produces a biometric that matches poorly to other biometrics of themselves.

These low match scores imply a higher than average false reject rate for Goats.

- Lambs: A person who is a Lamb can be easily impersonated. When the biometric of such a person is paired to a biometric from a different person the resulting match score will be higher than average. Consequently, false matches are more likely.

- Wolves: A person who is a Wolf is good at impersonation.When such a person presents a biometric for comparison they have an above average chance of generating a higher than average match score when compared to a stored biometric of a different person.

The distinction between a Lamb and a Wolf rests upon an assumed asymmetry between biometric signatures. In the face recognition nomenclature of target and query images, a target image is presumed to have been acquired in the past and is stored with the recognition system. It is compared to a query image, which is typically presumed to be a novel image acquired at the time recognition is being carried out.

In some face recognition contexts, the asymmetry makes sense, such as when query images acquired under uncontrolled lighting are compared to target images acquired under controlled lighting (FRGC Experiment 4 [5]). In other cases, where there is no meaningful distinction between query and target images, the Lambs and Wolves merge into a single category Lambs/Wolves.

Doddington et al. [2] provided precise definitions for Goats, Lambs and Wolves. To determine which people qualify as Goats on a particular dataset using a particular algorithm, first the average match score for each person is computed. Next, the people are sorted from lowest to highest match score. Finally, the first 2.5% of the people in this sorted list are labeled as Goats. These people have overall low match scores and are most likely to be falsely rejected during verification.

To identify Lambs, first the average non-match score for a person is computed from the pool of non-match scores where the person provides the target image. Next, the people are sorted from highest to lowest average non-match score. Finally, the first 2.5% of the people in this list are labeled as Lambs. The process for Wolves is nearly identical, with the change that average non-match scores are computed from all non-match scores where a person contributes the query image. When query and target images are interchangeable, a person is characterized by the average of all non-match scores for which they contribute an image and there is a single category Lambs/Wolves.

Sheep are defined by negation, all people not explicitly labeled as Goats, Lambs or Wolves are labeled Sheep. Most people are Sheep and as a category they are less interesting.

4. Evidence for Doddington's Zoo

To put our framework to the test, we select people, images, and similarity scores from FRVT 2006 experiments [3]; experiments 22 and 24. In experiment 22, still frontal face images taken with uncontrolled lighting are matched against still frontal face images taken with uncontrolled lighting. Experiment 24 matches target images taken under controlled lighting against query images taken under uncontrolled lighting. Experiment 22 is symmetric in so much as target and query images are interchangeable and consequently there is no meaningful distinction between Lambs and Wolves: we define a single category Lambs/Wolves. Experiment 24 is asymmetric and the distinction between Lambs and Wolves is maintained.

Testing for the existence of a first-order zoo requires a set of people and images such that the images can be randomly divided into two disjoint partitions. To further balance the tests, the same number of images are selected for each person. Finally, because images taken on the same day are more easily matched, no two images for a given person may be taken on the same day. Taking all these constraints into account, 16 controlled and 16 uncontrolled lighting images were selected for each of 257 people.

Because part of testing for the existence of a second-order zoo involves comparisons between algorithms, our analysis uses results for two top performing FRVT 2006 algorithms. These are two of the three algorithms used in the construction of the Good, the Bad and the Ugly Challenge Problem [4]. Tests for the existence of a first-order zoo will be run independently for both algorithms. The algorithms are designated as A and B.

4.1. Zeroth-Order Zoo

The zeroth-order Doddington's Zoo always exists by definition. In other words, people can always be sorted in terms of average match scores and the bottom 2.5% labeled Goats. Since all our testing will be over the same 257 people, there will always be 6 Goats. Likewise for Lambs and Wolves, we use the 2.5% cutoff originally proposed by Doddington. Hence, for the uncontrolled-to-uncontrolled (UU) scenario corresponding to Experiment 22 there will be 6 Lambs/Wolves and for the uncontrolled-to-controlled (UC) scenario corresponding to Experiment 24 there will be 6 Lambs and 6 Wolves. Henceforth experiments 22 and 24 will be referred by the scenarios UU and UC, respectively.

4.2. First-Order Zoo

The essential question upon which the existence of a first-order zoo rests is whether a person labeled as a Goat using data from a particular scenario is any more likely than others to be labeled a Goat using different data drawn from the same scenario. For the moment, let us concern ourselves just with Goats. The extension to Lambs and Wolves

is straightforward and will come next.

To make this notion of generalizing within a scenario more precise, a master data set is selected to characterize the scenario. Then, it will be randomly partitioned into two data sets and the number of people labeled Goats in both will be compared against the expected outcome if labeling was done independently for each data set.

Because of the symmetry of scenario UU and the asymmetry of scenario UC, the process of partitioning the master data set and labeling people is slightly different. For the UU scenario, 8 of the 16 uncontrolled images for each person are assigned to partition R_1 and the remaining 8 images are assigned to R_2. Then, for each partition and for each person, all combinations of 8 images provide match scores. Discarding images matching to themselves, the result is 56 match-scores per person.

Following Doddington's recommendation, each person is then characterized by the mean of these 56 match-scores. The 6 people with the lowest mean match-score are labeled as Goats in partition R_1. Likewise, the 6 people with the lowest mean match-score in R_2 are labeled as Goats.

At one extreme, if one believed that personal identity were the factor determining recognition difficulty, it would be expected that the same 6 people would be labeled Goats in R_1 and R_2. In other words, once a Goat, always a Goat. In practice, we do not observe anything approaching such an absolute, and we focus upon the frequency with which people are labeled Goats in both partitions.

For algorithm A and data set UU, the outcomes for 50 randomized trials is summarized in Table 1. By a randomized trial, we mean a new independent partitioning of the 16 images of a person into 8 for R_1 and 8 for R_2. To interpret, the first row indicates outcomes using the mean match score. Note that in 36 out of 50 trials, the intersection of the people labeled Goats in R_1 and in R_2 is of size zero. In other words, no person is labeled a Goat in both partitions. In 10 trials the intersection contains one person, and in only 4 trials are there 2 people in the intersection.

Table 1. Goats intersection counts using mean and median match scores. Column c0 is number of times the intersection is empty, c1 the number of times the intersection contains one person, etc.

measure	c0	c1	c2	c3
mean	36	10	4	0
median	24	20	2	4

4.2.1 Existence as a Statistical Hypothesis Test

The counts in Table 1 raise the key question "Is this outcome likely to happen by chance?". To answer this question, and formally define when a first-order zoo exists, we formulate two hypotheses: H0 and H1.

H0 The probability of a person being labeled a Goat in one partition is independent of their label in the other.

H1 A person labeled a Goat in one partition is more likely to be labeled a Goat in the other.

As is often the case with statistical hypothesis testing, it is easier to measure the probability of H0 given observations and reject H0 in favor of H1 when this probability, the p-value, drops below a threshold.

Under the null hypothesis H0, the number of people labeled as Goats in both partitions is governed by a hypergeometric distribution. To see this, consider the following connection to the more standard metaphor of drawing colored marbles from a jar without replacement. To begin, let us index people using integer identifiers 0 through 256 inclusive and further fix the labeling for R_1 such that people 0 through 5 inclusive are labeled Goats. Asking for the count of people labeled Goats in both partitions is equivalent to asking how many of the integers 0 through 5 are drawn at random and without replacement from the set of integers 0 through 256 inclusive. Finally, we can simplify this last statement to an equivalent statement, what is the probability of drawing o white marbles from a jar containing 6 white marbles and $257 - 6 = 251$ black marbles? The count of white marbles o is governed by the hypergeometric distribution, indeed this white and black marble example is commonly used to motivate the hypergeometric distribution.

The hypergeometric distribution for 6 Goats and 257 people total tells us the probability of zero and one people being Goats in both partitions is 0.8666 and 0.1268, respectively. Two or more people labeled as Goats in both partitions by chance is a rare event, with probability 0.0066. These probabilities can be mapped directly to expected counts for our 50 trial experiment: 43 and 6 expected cases with zero and one person labeled Goats in both, respectively, and 1 with 2 or more people labeled Goats in both. Intuitively, it is already apparent that the 4 out of 50 trials with 2 people labeled Goats is unlikely under H0.

To make this rigorous, a χ-square test between our observed overlap counts and those predicted by H0 is carried out and this provides the p-value for H0. In the case for row 1 of Table 1, the p-value is less than 10^{-4} and H0 is rejected by any standard threshold. Thus, for algorithm A and scenario UU there exists a first-order zoo for Goats.

4.2.2 Means vs. Medians

Row 2 of Table 1 shows the result using median match-score in place of mean match-score to characterize a person. The number of people labeled Goats in both partitions increases using the median, and thus a person labeled a Goat in one partition is somewhat more likely to be labeled a Goat in the other partition using the median. All of the tests carried out

Table 2. First order zoo tests for Goats, Lambs and Wolves. The p-value for H0 is less than 10^{-4} for all rows so we conclude that a first-order zoo exists for all animals and both algorithms. L / W represents Lambs / Wolves.

Animal	Algorithms	c0	c1	c2	c3	c4	c5
Goats	A,UU	24	20	2	4	0	0
Goats	A,UC	1	8	22	12	7	0
Goats	B,UU	18	26	6	0	0	0
Goats	B,UC	0	5	25	12	5	3
L / W	A,UU	14	21	11	4	0	0
Lambs	A,UC	0	1	14	18	12	5
Wolves	A,UC	1	11	17	18	3	0
L / W	B,UU	1	12	27	6	4	0

Table 3. Second order zoo χ^2 p-value to test the null hypothesis that the Goats and Lambs/wolves intersections come from the null distribution. The tests use the same algorithm with different class of underlying data, UU to UC.

Animal	Algorithm	c0	c1	χ^2 (p-value)
Goats	A	50	0	0.03495
Goats	B	46	4	0.465
Lambs	A	44	6	1
Lambs	B	33	17	0.0011
Wolves	A	48	2	0.1355
Wolves	B	50	0	0.03455

Table 4. Second order zoo χ^2 p-value to test the null hypothesis that the Goats and Lambs/wolves (L / W) intersections come from the null distribution. The tests use different algorithms with the same class of underlying data. Each scenario and animal combination is repeated, with the first indicating a comparison of algorithm A to B, and the next algorithm B to A.

Animal	Scen.	c0	c1	c2	c3	c4	χ^2 (p-value)
Goats	UU	40	9	1	0	0	0.2187
Goats	UU	46	4	0	0	0	0.4612
Goats	UC	2	19	19	8	2	$> 10^{-4}$
Goats	UC	2	21	18	9	0	$> 10^{-4}$
L / W	UU	44	6	0	0	0	1
L / W	UU	47	3	0	0	0	0.2681

for this paper have been done using both the mean and the median, and this result is typical. To give the zoo its best chance of generalizing, we therefore will adopt the median rather than the mean in all places where a set of match or non-match scores must be condensed to a single value.

4.2.3 Results for both Algorithms and both Scenarios

Table 2 summarizes the results of our tests for the existence of a first order zoo for both algorithms and both scenarios. The columns indicate the type of animal, the algorithm and the scenario, and then intersection counts over the 50 trials. More specifically, column c0 counts how often the intersection of people with the given animal name in the two partitions is empty, c1 counts how often one person is assigned the given animal name in both partitions, etc.

For the Lambs/Wolves results, all available non-match scores for each person are used within a partition in order to determine the median non-match score per person. This is true for both the UU and UC scenarios.

The construction of the random partitions for the UC scenario is slightly different because of the asymmetry between target and query images. For the UC scenario, the uncontrolled query images are selected from the same 16 used in the UU scenario, and divided at random into 8 for partition R_1 and 8 for partition R_2. The controlled target images are drawn from a new set of 16 controlled lighting images for each of the 257 people. There are 64 possible match-scores per person per partition in the UC scenario. To better match the analysis for the UU scenario, 56 of the 64 are selected at random when computing the median match score.

4.3. Second-Order Zoo

The existence of a second-order zoo is determined in a fashion similar to that defined for first-order zoo. As with the first-order zoo, we assign labels to people based upon two data sets and then consider the intersection of these two sets of people. However, the two data sets in this case involve either a change in algorithm, A versus B, or a change in scenario, UU to UC.

The null hypothesis H0 is the same, since it depends upon the simple idea that an animal label assigned to a person in one case is independent of the assignment of that animal label to a person in the other. Thus, p-values are still determined using the χ^2 test as already described.

Tables 3 and 4 summarize our findings with respect to the existence of second-order zoos for algorithms A and B and scenarios UU and UC. Table 3 keeps the algorithm constant while changing the scenario from UU to UC. Table 4 keeps the scenario the same while changing the algorithm, either from A to B or B to A. The columns c0, c1, etc., contain the intersection counts just as in Table 2.

Unlike the first-order zoos, whose existence was easily demonstrated, the existence of second-order zoos is problematic. In Table 3 using a standard cutoff of 0.05 for the p-value, a second-order zoo does not exist in 5 out of the 6 combinations of animals and algorithms, with the first row and the last failing to reject the null hypothesis owing to the two-tailed characteristic of the hypothesis test. A conservative cutoff of 0.01 also maintains 5 out of 6 rejections of the

existence of the second order zoo. In Table 4, the existence of a second-order zoo is supported only for 2 of the 6 cases.

4.4. Third-Order Zoo

For a third order zoo to exist, the likelihood a person is assigned a label using one algorithm and one scenario must be predictive of the person receiving the same label using a different algorithm and scenario. Again, the H0 hypothesis test is the same as before, depending solely upon the intersection counts. In all our tests, the intersection counts were too low to support the existence of a third-order zoo.

5. Yager and Dunstone's Zoo

Doddington's definitions capture something clearly of interest, but fail to capture relationships between match and non-match scores. To illustrate, a person with universally low match and non-match scores will qualify as a Goat, even though their low non-match scores mean they are actually in no danger of being falsely rejected. This concern for capturing relationships between match and non-match scores brings us to the work of Yager and Dunstone [9]

5.1. Doves, Chameleons, Phantoms and Worms

Doves, chameleons, phantoms and worms are defined by Yager and Dunstone [9] in terms of extreme match and non-match scores and more importantly relationships between them. As was done above, median match scores are computed for each person, and then the people are sorted from lowest to highest median match score. Similarly, a separate sorted list of people is created where people are sorted from lowest to highest median non-match scores.

The approach of Yager and Dunstone now differs from that of Doddington et al. in that intermediate labels are assigned based upon the top and bottom quartiles in the sorted lists just defined. Specifically, people in the group $\mathcal{M}_h$ are the 25% of the people with the highest average match scores. The people in the group $\mathcal{M}_l$ are the 25% of the people with the lowest average match scores. Analogous sets exist for the non-match scores, where the people in the group $\mathcal{N}_h$ are the 25% of the people with the highest median non-match scores, and the people in $\mathcal{N}_l$ are the 25% of the people with the lowest median non-match scores.

Yager and Dunstone's animals are defined by the intersections between combinations of these four groups:

- Doves: A person who is a dove matches very well against themselves and poorly against others. They are easily recognized and are defined by the set $\mathcal{M}_h \cap \mathcal{N}_l$.
- Chameleons: A person who is a chameleon matches well in general, both to themselves and to others. They are likely to cause false accepts but not false rejects. Chameleons belong to the set $\mathcal{M}_h \cap \mathcal{N}_h$.
- Phantoms: A person who is a phantom matches poorly in general, both to themselves and to others. They are likely to cause false rejects but not false accepts. Phantoms belong to the set $\mathcal{M}_l \cap \mathcal{N}_l$.
- Worms: A person who is a worm matches themselves poorly and other people relatively well. They result in disproportionate number of errors, both false rejects and false accepts. Worms belong to the set $\mathcal{M}_l \cap \mathcal{N}_h$.

Although Yager and Dunstone concluded that Goats, Lambs and wolves are common in biometric systems, the reasons for the existence of a particular animal group are complex and varied. According to them, people are not inherently unsuitable for biometric identification and this notion has been highly exaggerated. Factors accounting for these animal groups include enrollment procedures, feature extraction and match algorithms, data quality, and intrinsic properties of the user population. Yager and Dunstone concluded that "Doddington's Zoo phenomena may be image specific as opposed to individual specific".

6. Evidence for Yager and Dunstone's Zoo

For simplicity, here we will consider only the uncontrolled to uncontrolled (UU) scenario, and our categorization of people as Doves, Chamelons, Phantoms and Worms will be based upon the same 257 people and 16 images of each person randomly divided into partitions R_1 and R_2. Here too we will replace the use of means with medians.

6.1. Zeroth-Order Zoo

Because Yager and Dunstone's animals are defined by the intersections of sets, the existence of a zeroth-order zoo is not guaranteed. While unlikely, it is possible that a data set may fail to contain any instances of an animal. This caution aside, in our experiments we see variation in the number of people labeled as different animals, but there are always examples of all types of animals.

Figure 1 shows histograms of the number of Doves, Chameleons, Phantoms and Worms found in our 100 partitions. There 100 partitions because we have two partitions for each of the 50 trials. Note that the range of variation in the number of animals changes between algorithms. For example, algorithm B produces more Worms than Doves while A produces more Doves than Worms.

6.2. First-Order Zoo

The results for our tests of existence for a first-order Yager and Dunstone zoo are summarized in Table 5. What we observe is that for both algorithms a first-order zoo exists for Doves, but not for any of the other animals. This is interesting in several respects, not least of which because Doves are the most desirable of the animals.

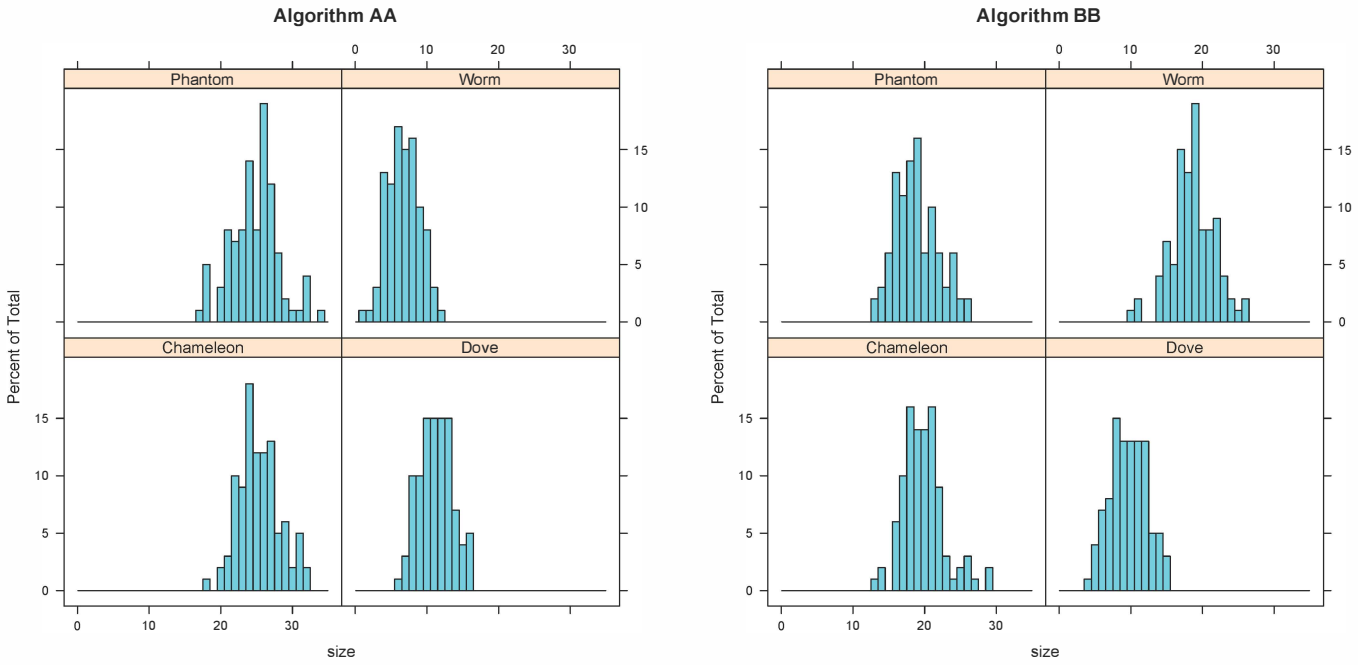

Figure 1. Histograms showing size of the sets of people labeled as Doves, Chameleons, Phantoms and Worms for algorithms A and B.

Table 5. First order zoo χ^2 p-value to test the null hypothesis that the chameleons, phantoms, doves and worms intersections come from the null distribution.

Animals	Algorithms	χ^2 (p-value)
Doves	A	0.0134
Chameleons	A	0.3605
Phantoms	A	0.6527
Worms	A	0.0377
Doves	B	0.0104
Chameleons	B	0.1141
Phantoms	B	0.08455
Worms	B	0.05655

One detail concerning how the p-values are obtained for these tests deserves mention. The use of the hypergeometric distribution for Doddington's zoo was relatively straight forward because the number of animals remained constant as a fixed percent of the total number of people. As already highlighted in Figure 1, this is not true for the Yager and Dunstone zoo. Therefore, in order to determine the expected intersection counts for these animals, it is necessary to take into account the fact that when there are more people labeled as an animal, the likelihood of having more people in the intersection goes up.

To illustrate how this was handled, consider Doves for algorithm A. A bootstrap process was used to sample pairs of set sizes from the 100 observed number of Doves shown in the histogram. This provides us an estimate of the probability of comparing say 10 Doves to 12 Doves or 11 Doves to 5 Doves. For any specific pair of set sizes, the hypergeometric distribution defines the probability of a zero intersection, an intersection of size one, etc. The probability density function we seek is therefore found by integrating over the different set sizes, and since the space of sets sizes is discrete, integration becomes summation.

6.3. Second-Order Zoo

We also tested for the existence of any second-order zoos by comparing across algorithms and the results are shown in Table 6. For Doves, Chameleons and Phantoms the p-values are too high for us to reject H0 and we must conclude there is no second-order zoo. For Worms, the p-values fall below a standard 0.05 cutoff. However, a look at the intersections reveals this is because there are more zero intersections than H0 would predict. Thus, there is actually an inverse relationship - being a Worm in one partition lessons the change a person is a Worm in the other. It would be unwise to overplay this last result. Suffice it to say, the p-value is not evidence for a second-order Worm zoo. Given the lack of any second-order zoos, there is no point in looking for evidence to support the existence of a third-order zoo.

7. Summary

The presence or absence of a biometric zoo is not a straightforward narrative and often depends on the framework. In this paper we have defined and presented results

Table 6. Second order zoo χ^2 p-value to test the null hypothesis that the chameleons, phantoms, doves and worms intersections come from the null distribution. The tests use either the same algorithm with different class of underlying data or different algorithm with the same class of underlying data.

Animal	Algorithms	χ^2 (p-value)
Chameleons	A - B	0.4219
Phantoms	A - B	0.2684
Doves	A - B	0.1802
Worms	A - B	0.0372
Chameleons	B - A	0.4145
Phantoms	B - A	0.2652
Doves	B - A	0.07085
Worms	B - A	0.03525

for the zeroth, first, second and third order zoo frameworks. These frameworks and the results presented in this paper can be better illustrated through Figure 2. The representative numbers are the mean intersections of the goats for the corresponding algorithm and the scenario. Each circle represents a first order zoo. The horizontal and the vertical lines represent the second order zoo and the diagonals are the third order zoo, respectively. In the context

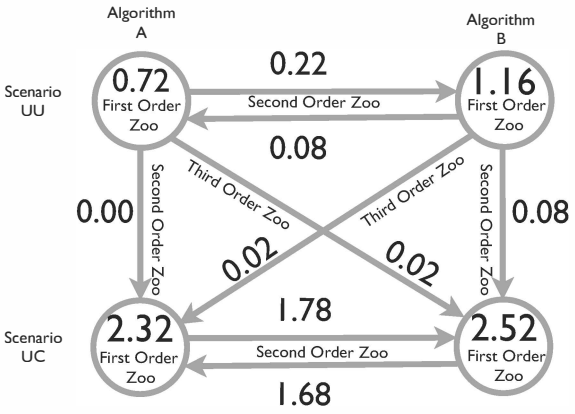

Figure 2. Illustration of the zoo framework using mean intersections for Goats.

of this figure, each of these frameworks can be further illustrated as follows. For the first order zoo consider for example top left circle. It is the mean intersections when algorithm A is used with UC scenario. In this combination out of total 50 trials, 24 show no overlaps, 20 have atleast 1 common goat between the two partitions of a trial, 2 have atleast 2 common goats and 4 trials have three goats common between the partitions. This implies that there are $\frac{24\times0+20\times1+2\times2+4\times3}{50} = 0.72$ mean intersections and hence the number within the circle. We can similarly find the mean intersection numbers for rest of the circles.

The second order zoo represented by the horizontal and the vertical lines can be explained as follows. The top horizontal line with a mean intersection of 0.22 shows a framework where we use cross algorithms A - B and the scenario UC. The horizontal arrows differ from the vertical arrows because in the former case we use the same scenario but different algorithms and in the latter case same algorithm with different scenarios. These lines represent the results of Tables 3 and 4, respectively. The third order zoo represented by the diagonal lines shows mean intersection values when different algorithms A and B are used with different scenarios, UU and UC.

We found a strong evidence for the presence of a first order zoo in Doddington animals. However, majority of the cases (9 out of 12) in the second order zoo show contrary evidence. In addition, Yager and Dunstone menagerie does not exist in the majority of the cases (6 out of 8) in the first or the second order zoos. There is no evidence of the presence of the third order zoo in either Doddington animals or the Yager and Dunstone menagerie. Therefore it is safe to conclude that zoos of order greater than first are either scarce or totally absent.

References

[1] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, 113(6):750–762, June 2009.

[2] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *Proc. ICSLD*, 1998.

[3] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. Frvt 2006 and ice 2006 large-scale results. Technical report, NISTIR 7408, NIST, 2007.

[4] P. J. Phillips, J. R. . Beveridge, B. Draper, G. H. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. A. Sahibzada, and S. Weimer. An introduction to the good, the bad, & the ugly face recognition challenge problem. In *FG*, 2011.

[5] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *CVPR*, 2005.

[6] N. Poh, S. Bengio, and A. Ross. Revisiting Doddington's Zoo: A Systematic Method to Assess User-Dependent Variabilities. In *MMUA*, 2006.

[7] N. Poh and J. Kittler. A Biometric Menagerie Index for Characterising Template/Model-Specific Variation. In *ICB*, 2009.

[8] A. Ross, A. Rattani, and M. Tistarelli. Exploiting the "Doddington Zoo" Effect in Biometric Fusion. In *BTAS*, 2009.

[9] N. Yager and T. Dunstone. The biometric menagerie. *PAMI*, 32(2):220–230, 2010.