On the Paper of Witztum, Rips and Rosenberg on Equidistant Letter Sequences in the Book of Genesis

Gil Kalai *

December 10, 1997

Abstract

In their paper [12] Witztum, Rips and Rosenberg (WRR) claimed to have statistically proved the existence of a hidden code in the Book of Genesis giving information about personalities living many centuries after the Book of Genesis was written. Their test was first conducted on 34 most prominent Jewish figures and another test was made on a second list of 32.

We give statistical evidence that the work of WRR is deceptive as well as interesting insight into the nature of their flaws. Our findings suggest that there was some optimization process in choosing the data for WRR's experiments.

Specifically, our findings indicate that the principal statistics (the P2-statistics) used by WRR played a crucial role in the optimization itself for both experiments and in the stopping rule for the optimization in the second experiment.

Our work complements the comprehensive study by Bar Hillel, Bar-Natan and McKay who studied various aspects of the two experiments conducted by WRR and reached the same conclusion.

*Institute of Mathematics, Hebrew University, Jerusalem 91904 Israel.

1 Introduction

1.1 The paper of Witztum, Rips and Rosenberg

This paper contains some findings concerning the work [12] of Doron Witztum, Ilya Rips and Yoav Rosenberg (briefly WRR) in which they claimed to have statistically proven the existence of a hidden code in the Book of Genesis containing information about personalities living many centuries after the Book of Genesis was written. They claim to have shown significant proximity of Equidistant Letter Sequences (briefly ELS) of names and appellations of certain Rabbis w.r.t their (known) dates of death and birth. Their first test was conducted on a list of the most prominent Jewish figures and the results were impressive. WRR were asked to repeat the experiment on a second list of Rabbis and the results were *very similar*.

The first list of 34 Rabbis consists of all Rabbis whose entries in the *Encyclopedia of Great Men in Israel* by M. Margalioth is at least 3 columns long (such that either their date of birth or date of death is mentioned there). The second list consists of Rabbis having 1.5-3 columns in that encyclopedia.

WRR developed a statistical way to associate to a pair of terms a and b a certain "distance" $c(a,b)$. (We call the values of $c(a,b)$ - *pair distances.*) Thinking about a and b as two "massages" hidden in the text as equidistant letter sequences $c(a,b)$ describe how close these "messages" are. $c(a,b)$ is a rational number between 0 and 1 which can be regarded as a probability describing how lucky we are that a and b are found so close together as equidistant letter sequences in the Book of Genesis (We regard the function $c(a,b)$ as a black box, for their precise definition see [12, 11]). As an example, WRR compute in [11] the value $c(a,b)$ where a and b are two words related to the Jewish holiday *Hanukka.* The value $c(a,b) = \frac{2}{121}$ is indeed a small number.

WRR considered various appellations for each Rabbi and various standard forms for writing his known dates of birth and/or death. They computed all the pair-distances between all pairs of terms consisting of one appellation and one date for the same Rabbi. For example, Rabbi number 1 in the second list has five different appellations four of which appear in Genesis as ELS's, and his date of death can be written in three forms which are all presented as ELS's in Genesis. This

Rabbi thus contributes 12 pair-distances to the second experiment.

WRR obtained 152 values for the first list and 163 values for the second. If these pairs were not hidden in the text, they (implicitly) argued, then the numbers we would see will be uniformly distributed in the unit interval. However, if there is a hidden text containing the appellations and dates of the Rabbis, then we would see many small pair-distances.

WRR developed some statistical measures to check if the set of pair-distances is not uniform according to a hidden text assumption. They computed the significance obtained from these statistical tests for the combined list of numbers for all the Rabbis, and found (in both experiments) extremely high significance levels. (For more details, see [12] and [11].)

The principal statistical measure used by WRR is the "P2-statistics" - the probability that the product of uniformly distributed independent random variables is smaller than the product of the numbers observed in the experiment. (See Section 7.1.)

1.2 The ground rules for this work

1. We do not enter at all to the historical, grammatical or biographical decisions made by WRR.

2. We do not make any new computation of the function $c(w, w')$ on the book of Genesis or on any other text. We just use the two sets of 152 numbers and 163 numbers obtained by WRR in the first and second experiments respectively. A file containing this data (and explanation) can be found at [1].

3. Findings supporting the hypothesis of optimization are considered only if they cannot be explained (as far as we know) even by WRR's original hypothesis of hidden code in the text of Genesis.

1.3 Our findings

We give statistical evidence for the hypothesis that the significance in the second test of Witztum, Rips and Rosenberg is the result of an optimization process in choosing the data, which was stopped when the significance level (given by the P2-statistics) of the first test was reached.

Our results further suggest that the optimization process was carried out (at least in its final stages) by adding or deleting favorable appellations for the Rabbis until the addition or deletion of the last appellation moved the significance level beyond that of the first test. It goes without saying that such a procedure, which consists of manipulating the data to reach the desired goal, is completely illegitimate.

We give further statistical evidence suggesting that there was a similar optimization process for choosing the data in the two experiments. We indicate some statistical dependence between the data of the two experiments which is completely unexpected.

The basic idea behind both these findings is similar, we show unreasonable proximity between the significance levels of the two experiments and between the distributions of pair-distances of the two experiments.

Finally, we consider the distributions of pair-distances in each of the experiments. We show that these distributions do not support reasonable interpretations of the original research hypothesis of a hidden text. On the other hand these distributions are close to distributions obtained by simple simulations of an optimization procedure based on the P2 statistics.

Our findings complement those of Bar Hillel, Bar-Natan and McKay, see Section 10.

Part I: Similarity of the P2-scores

2 Some Examples

The case in hand is complicated and emotionally loaded. Moreover, the (hidden) assumption of divine intervention complicates things even further. Let us start, therefore, with simple examples which demonstrates the basic approach.

Suppose someone claims to be able to hit a globe hanging 200 meters away with a bow and arrows while blindfolded. You blindfold him, he shoots once, then after a while he shoots again, and then sends his son to bring the globe, and ...lo and behold! Both arrows are stuck in the globe on the equator, very close to each other. What the boy did not know, however, is that while his father was shooting, the globe was spinning round the axis through its poles very quickly, so that if an arrow hits the globe the longitude on which it lands is essentially random. Taking a closer look you calculate that the probability of two random points on the equator being as close as the two arrows is 1/100. Furthermore, the distance between the two arrows is about the closest the boy could have stuck the arrows without hurting his fingers if he was the one who stuck them in....

We observe here a phenomenon which supports a simple cheating strategy and is very unlikely to be found otherwise.

If we also realize that the angles in which the arrows are stuck do not fit to the claim that they were shot from distance but instead are compatible to the height of the son with respect to the position of the ball this will give us additional support for the cheating hypothesis.

We will move now to an example which is quite close to the discussion in the next sections.

A researcher conducts a statistical test to check the hypothesis that there is a significant positive correlation between height and salary among people with academic education.

She instructs her assistant to interview all people from one neighborhood in the city and to sort out the 90 with academic education among them.

Then she tests for each individual his height and his salary. She finds indeed a positive correlation and the significance level for her finding is 0.000132.

When she is asked by the referees of the paper to repeat her experiment she assigns again her assistant to the job. He chooses another neighborhood, this time it is a neighborhood with many immigrants so the criteria for people to be regarded as having academic education is harder. The assistants repeats the process (finds 80 persons this time) and proves again that there is positive correlation between height and salary with a significance level of 0.000120.

At this point the professor suspects that the assistant who wanted to ingratiate himself with her faked the results using the degrees of freedom he had. In particular, his precise criteria for a person to be qualified as having academic education could not be understood by her.

The similarity of the two significance levels raises the hypothesis that however the assistant tampered with the data his strategy was to do it gradually until the significance level of the second test passed that of the first test. The ratio between them is 1.1.

How can we test such a hypothesis? Assume that we have all the relevant data on the people considered as having academic education but not on those rejected.

Suppose we split the 170 chosen individuals in another way into two groups of people consisting of 90 and 80 people respectively. We can assume the significant positive correlation between height and salary will hold also for both these groups. However, there is no reason to assume that the proximity between the significance levels we observe in the two parts will typically be higher then that we observe in the original research. In fact, there are reasons to assume it will be typically smaller.

We can bound the significance of the proximity of the two significance levels by comparing it in a Monte Carlo experiment to the proximity observed by splitting the 170 chosen individuals to 90 and 80 at random.

Suppose we find out that with probability 1/100 the following event occur: *For a random splitting of these 170 people into two parts of 90 and 80, the ratio of the significance levels for the two parts is 1.1 or smaller.*

This would be an incriminating evidence since there is here a phenomenon that we could expect to occur if the assistant was tampering with the data in a certain way but was very unlikely to be found otherwise.

But we can make one further step. Suppose we suspect that the main degree of freedom of the assistant was in including or rejecting persons in his list and that he gradually added people with academic education to the second list which were favorable to the research hypothesis until he reached the significance level of the first test. In this case the proximity of the significance level of the two tests should be related to the effect of adding the *last person*. If the typical effect of adding a single person to the list is compatible with the number 1.1, this will give an additional support to the cheating hypothesis. (As we will see later, being compatible means that typically adding a person with academic education to the list changes the significance levels by a factor in the neighborhood of 1.1^2.)

If the researcher will discover unexplained statistical dependence between the data of the two experiments this could further support the hypothesis that the results cannot be trusted.

Several people were reminded by the Bible code case of the following story (possibly a tale), which I first heard from Maya Bar Hillel. The mathematician Poincaré bought loaves of bread from a certain bakery and after a while he complained to the police that the average weight he observed is 0.9 kilograms rather than the required 1 kilogram. The police intervened and since then all of Poincaré's loaves of bread were heavier than 1 kilogram. Six months later Poincaré was asked if the baker stopped cheating and his answer was that he didn't. He found statistical evidence to the fact that the baker kept cheating but that every day he was putting aside for Poincaré a loaf of bread which was heavier than 1 kilogram. Although all the loaves of bread Poincaré got were over 1 kilogram their distribution was significantly close to a normal distribution with average 0.9 kilogram cuts off at 1 kilogram.

Poincaré identified a simple cheating strategy of the baker so that the distribution of weights of his loaves of breads agrees with the distribution you expect if you assume cheating and is very unlikely to happened if there was no cheating.

All the examples of this section have the weakness that the statistical analysis was not made to give a priori predictions on new data but rather to study given data. It is important to note, however, that in all our examples there were strong a priori reasons to suspect that some deception is taking place. With a good lawyer the baker may get off the hook. But the researcher, as soon as she got the picture, disqualified the experiments made by her assistant.

3 The significance levels in the research by Witztum, Rips and Rosenberg

Witztum, Rips and Rosenberg claim to have found an arrow sent by the author of the Book of Genesis, crossing thousands of years in its flight.

In their 1987 preprint [11] which presented the situation after the second test was carried out, Witztum, Rips and Rosenberg write in the introduction as follows:

"For the string G however, for the unperturbed sample we obtained $c(w, w')$ tending to zero with a probability against a null hypothesis of a uniform distribution that we estimate as $1.3 \cdot 10^{-9}$ for the first experiment and $1.2 \cdot 10^{-9}$ for the second (which gives the probability $1.8 \cdot 10^{-17}$ for the union of the samples)."

These numbers refer to the principal measure of significance used by WRR at that time for the two tests they made. The measure of significance used was later called the P2-statistics or the P2-score. In the Analysis Section of the same preprint the next digit is revealed so the numbers are $1.29 \cdot 10^{-9}$ and $1.15 \cdot 10^{-9}$, respectively. So in simple words what they are saying is that the significance level of their first experiment is $1.29 \cdot 10^{-9}$ and that of the second experiment is $1.15 \cdot 10^{-9}$.

The ratio between these two numbers is 1.1217. We will see that in view of the instability of the P2-statistics this ratio is extremely small. It is significantly small compared to random partitions of the 66 Rabbis into sets of 34 and 32 Rabbis respectively, it is significantly small compared to random perturbations of the original division of Rabbis, it is even substantially smaller than the typical effect of adding and deleting a *single* appellation to a single Rabbi.

The only explanation we can offer for this phenomenon is that there was an optimization process in the second experiment which stopped when the significance level of the first experiment was reached. Moreover, consider an optimization process in the second test which terminates with the addition of an appellation which brings the significance level beyond that of the first test. We can expect that the ratio of the significance levels of the two tests will be in the neighborhood of the square root of the effect of the last appellation. Indeed, the square root of the typical effect of adding an appellation is comparable to the ratio we witness.

It is worth noting that the astronomical significance levels cited above (from the 87 preprint [11]) are false due to wrong independence assumptions.

4 Random partitions

I considered a random partition of the 66 Rabbis from the two tests together into one part of 34 Rabbis and another part of 32 Rabbis, and studied the distribution of the ratios of the P2 scores for the two groups and, in particular, how likely it is that such a ratio is smaller than 1.1217, the ratio of the numbers reported by WRR.

The distribution of the ratio of P2 statistics is quite interesting. The probability that this ratio is below 1.1217 is roughly 1/100. (More precisely it is 0.0092, but see the technical remark in Section 7.) Note that this is a direct Monte Carlo estimate and it does not rely on any probabilistic assumptions. The median value of the P2-ratio is roughly 700. The average is huge due to rare occurrences of very high ratios and I cannot estimate it. The average of the logarithm (with base 10) is roughly 3.3 .

As an illustration, if you move Rabbi number 7 (Rabbi David Ganz) from the first to the second list (and it is agreed that he was on the first list by mistake (see [1]) according the criteria of WRR,) the ratio in question changes to 4.1268. As another illustration, the ratio we observe will be obtain by multiplying one out of the 152 pair-distances of the first experiment by 0.66, and leaving all other 151 pair-distances as they are.

Remark: We compute the probability that the ratio between the two P2-scores will be smaller than 1.1217. Of course, if we restrict our attention only to cases where the P2-score of the second test is smaller than that of the first test this further reduces the probability roughly by a factor of two.

5 The rationale behind the experiment and the basic consequences

We cannot expect the partition made by WRR to behave like a random partition. We should therefore explain what is the rationale for comparing its P2 score to a random partition. I will discuss separately two possibilities. The first is that the significant phenomenon described in the paper was the result of some optimization in choosing the data and the second is the original research hypothesis of WRR.

5.1 Assuming optimization took place

The explanation of WRR's results by optimization was suggested by various people and there is a comprehensive ongoing study by Bar Hillel, Bar-Natan and McKay concerning this possibility. It was previously believed that in both experiments there was some optimization (either intentionally or unintentionally) in order to improve the significance as much as possible. The above experiment suggests that, in fact, in the second test the results were improved until reaching the level of significance of the first test.

Indeed, without assuming this stopping procedure, there is no reason to believe that the P2 ratio of the original partition will not be smaller than that of a random partition and in fact, there are reasons to believe the contrary: in the two tests we have two distinct populations of Rabbis, the possible optimizations are different in the two tests (some parameters were determined by the first test) and the optimization skills are perhaps improved between them. For the random partitions we blend Rabbis from both lists. We can therefore assume that while each of the two parts will be less homogeneous they will be closer to each other than the original partitions.

5.2 Assuming the original research hypothesis

Let's go now to the original hypothesis of WRR. Here too, not only is there no reason to believe that the two P2 scores will be more proximal to each other than those of a random partition but again there are reasons to believe the contrary.

Since one list is of the more important Rabbis and the other is of the less important ones, we can expect that the knowledge of the historical data will be different and also that the hidden biblical code will treat them in a different way. Thus, the two populations of Rabbis in the original partition are quite distinct, but in a random partition, where in each part we blend Rabbis from the two original lists, we can expect the two parts will be closer together.

5.3 Comparing with random perturbations of the original partition

The P2 ratio is not only significantly small with respect to random partitions but also with respect to random perturbations of the original partition. Thus, when you randomly replace k Rabbis on the first list with k Rabbis on the second list the probabilities of getting the P2-ratio of WRR or a smaller ratio are for $k = 1, 2, 3, 4, 5$ respectively: 0.035906, 0.022370 ,0.017372, 0.015322, 0.014150.

6 The effect of one appellation and a closer look at the stopping rule

One can say more on the possible stopping rule for the second test and where the P2 ratio we see could come from. This is related to the effect of adding (or deleting) of *one* appellation. (Remember, each Rabbi has several appellations.)

The ratio of the P2 scores is still considerably lower than the usual effect on the P2 scores obtained by a single addition of a single appellation. But the *square* of the ratio obtained by adding one appellation seems quite close (still a bit on the lower side) to the ratio we experience.

Indeed in the second list there are 44 appellations whose deletion decreases the P2 score, 23 whose deletion increase the P2 score and 35 appellations which are dummies. (The "dummies" do not participate in any pairs of ELS and therefore they have no effect on the score.)

For 12 out of the 44 appellations whose deletion decreases the P2 score, the amount of decrease is smaller than the square of 1.1217 (roughly 27 %). This is the case for 9 out of the 23 appellations whose deletion increases the score (39 %).

This is compatible with an optimization process of repeated additions of appellations which stops when reaching a score which passes the P2 score of the first test. Indeed, in such an optimization process the ratio between the resulting P2 score and the P2 score of the first test is likely to be in the neighborhood of the square root of the effect of adding the *last* appellation. to the P2-score.

To see this pass to the logarithm of the P2-score and note that when we gradually add quantities and stop when we pass a threshold T then we can expect the difference between the outcome and T be in the neighborhood of the half the last quantity that was added.

Alternatively, the ratio we witness is compatible to an optimization process where among a large pool of possible appellations the appellations whose deletion increase the P2-score are repeatedly deleted until a score which passes the original P2-score is reached.

We cannot tell if such an optimization was blunt cheating or if there was an innocent process (but totally wrong, of course,) where the criteria for including or rejecting appellations were formed using the pair-distances observed in the Book of Genesis.

The realization of the large degree of freedoms concerning the appellations by Dror Bar-Natan (et al.) was a major turning point in understanding the work of Witztum, Rips and Rosenberg.

Bar-Natan [1] discovered that the P2-score of the dates of the Rabbis in the second list with respect to the appellations of these Rabbis which were *not chosen* by WRR is significantly smaller than the the P2 score of the same dates with the same appellations randomly permuted.

In the example we gave in Section 2 Bar-Natan's finding is analogous to finding a significant *negative* correlation between height and salary for people who could have been considered as having academic education by the assistant but nevertheless were rejected.

7 Some explanations on the statistics and significance levels in WRR's work

7.1 The P2 statistics

The P2-statistics describes the probability that the product of n *independent* random variables distributed uniformly in the interval $[0, 1]$ will be smaller than a number x. It is given by (see [12])

$$F(n, x) = x(1 + y + \frac{y^2}{2} + \frac{y^3}{3!} + \cdots + \frac{y^{n-1}}{(n-1)!}), \qquad (1)$$

where $y = -\log x$.

7.2 The P1 statistics

In the analysis section of [11] WRR use also another statistics - later called the P1-statistics. The P1-statistics is the difference, in terms of standard deviations, between the the number of pair-distances whose value is smaller or equal than 0.2 and 0.2 times the total number of pair-distances. (In [12] WRR used the precise formula for the probability that the number of pair-distances ≤ 0.2 is as large as observed in the experiment assuming the pair-distances are uniform.)

The P1 statistic reflects the belief that because of the "hidden code" there will be many small pair-distances. P2 replaced P1 as the principal measure of significance because WRR felt that the

P1 statistics treats all pair-distances smaller than 0.2 in the same way while they expected many pair-distances to be much smaller.

7.3 The pair distances

The program computing the pair-distances using the algorithm of WRR had repeated minor modifications (or debugging) and we do not have now the program used for the numbers appearing in the Statistical Science article and not in the earlier preprints. We used the pair-distances which appear in [11] (correcting a single typo with respect to the earlier 86 preprint.) This data extracted from [11] can be found in [1]. The P2 scores we obtain from the pair-distances we collected from the preprint is close but not identical to the reported scores. We got $1.241456 \cdot 10^{-9}$ for the first experiment and $1.096396 \cdot 10^{-9}$ for the second. The ratio of these numbers is 1.1323. The probability to be below the ratio of the reported P2-scores is 0.0092 and the probability to be below the ratio we computed is 0.0010.

7.4 What is the true significance level of WRR's research?

The astronomical significance levels from the 87 preprints [11] are false (mainly) due to wrong independence assumptions. The pair distance $c(w, w')$ can be regarded as a function of the shortest ELS for w, the shortest ELS for w' and the distance between them. (The same remark applies to P1.) This creates massive dependencies among the pair distances. The continued use of false significance measures in the bible code activity is one of the reasons for the self-deception which is so characteristic of that activity.

The significance level reported in the paper [12] used a basic idea suggested by Diaconis and is $1.6 \cdot 10^{-5}$. The significance level for the second test if one implements the Diaconis test precisely as suggested is around $1.6 \cdot 10^{-3}$. Since the phenomenon claimed in [12] is vague it is hard to judge what is a correct statistical measure for its significance but there is no dispute that the pair-distance distributions exhibited by WRR are not random.

7.5 The significance of the significance level

WRR and others regard the high significance levels as some sort of a probabilistic proof of the claims made (or implied) in their paper. This by itself is based on a misunderstanding. While the significance level gives you (some) indication on the possibility that the results came by chance they do not help against various types of systematic errors and do not cover for absent or wrong interpretation.

Part II: The pair-distance distributions

In the heart of each of the two experiments of WRR were the pair-distances for all the pairs of appellations versus dates. There are 152 pairs for which the distances are defined in the first experiment and 163 in the second.

The pair-distance histograms namely the histograms of the pair-distances occurring at each experiment, are described in [12],p. 437 and [11]p. 4,5.

8 Similarity of the pair-distances distributions

As mentioned above, an unexplained similarity between the two experiments is a bad sign. We will consider now the similarity of the distributions of pair-distances between the two experiments.

The common measure of distance between two distributions is the supremum norm of their difference (we refer to this distance as the *sup-norm*-distance. See Chapter 14 of [14]. This can be visualized as follows:

Given two lists A and B of numbers (say, lists of the same size), order their union and define a random walk as follows: look at the i-th element and go right if the element is from A and left if it is from B. When you consider pairs of samples from the same distribution then the distribution of walks you will obtain by this process will be just the distribution of the ordinary random walk. If the two lists come from sampling different distributions then this walk will deviate more from the origin than an ordinary random walk. This is behind the Kolmogorov-Smirnov statistics for similarity of distributions.

We studied by a Monte Carlo experiment how significant is the proximity between the pair-distances distributions for the original partition of Rabbis compared to random partitions.

8.1 How similar are the distributions?

The sup-norm distance between the two distributions of pair-distances is 0.05489. See Figure 1(a) for the two distributions. The probability that for a random partition of the 66 Rabbis into parts of 34 and 32 the sup-norm distance between the two distributions of pair-distances is smaller or equal 0.05489 is around 0.035 .

We could expect the opposite phenomenon because for a random partition of Rabbis in both parts we blend the pair-distances of the two experiments.

The probability for a random partition to have a smaller sup-norm distance of distributions *and* a smaller ratio of P2-scores than the original partition is 0.0006. Thus, these two parameters strongly distinguish the original partition of Rabbis in a way which is contrary to what can be expected. The proximity of the pair-distances distributions has a positive correlation with the similarity of the P2-scores discussed earlier.

We also compared the sup-norm distance between the two original distributions to the sup-norm distance obtained by partitioning all the 263 pair-distances from the two experiments combined at random into two parts of sizes 152 and 163 respectively. The probability for a smaller sup-norm distance than 0.05489 is around 0.051. This corresponds to the sup-norm distance between two random samples of the same distribution which is the distribution of the pair-distances of the two experiments combined.

8.2 Possible interpretations:

A-priori there is no reason to believe that the two sets of pair-distances coming from the two experiments of WRR will behave like two samples of the same distribution. There are reasons to believe the contrary. However, as it turns out the sup-norm distance between these two lists of pair-distances is significantly small *even* for two samples of the same distribution. This indicates some dependence of the data from the two experiments. What can the reason for such a dependence be?

Here is an example: consider two institutes for mediocre research in two parts of the world, which consider two candidates every year. In order to keep their renowned mediocrity level they hire both candidates if and only if one is above the average and one is below the average, otherwise they hire none of them. Suppose the levels of the candidates are independent random variables uniformly distributed in [0,1]. The common hiring strategy will create dependency among the distributions of the levels of the faculty members of these two institutions.

The sup-norm distance between the distributions of levels of the faculty members in the two institutes will be smaller than the distance between distributions obtained by taking all the faculty members in the two institutions and dividing them at random. The same will apply to two institutes of shallow research which hire both candidates if and only if at least one of them is below the average.

The only explanation I have for the strong proximity between the two pair-distances distributions (if it didn't happen by chance) is a similar optimization process applied to both. Like in the hiring process in the example above, each decision in the optimization process consists of accepting or rejecting a bunch of pair-distances, like the pair-distances corresponding to one appellation w.r.t. the various ways to write the corresponding dates. However, I *do not* have an explanation for this strong proximity based on a specific optimization process in the case at question.

8.3 Further remarks:

1. The fact that the data described by the two WRR experiments seems to represent the same distribution (or two very similar distributions) suggests that this distribution is meaningful either in connection with the phenomenon WRR are claiming or in connection with an optimization process leading to their results.

However, this fact by itself is disturbing to the integrity of WRR's paper if we take into account other facts and other claims made by WRR. This robustness of the pair-distances distributions seems contradicting to

a. The big instability of the phenomenon claimed by WRR - there is a big difference how different Rabbis contribute to the phenomenon. The phenomenon is supported to a large extent on a small number of Rabbis.

b. The noise added by wrong or unexplained decisions made by WRR.

c. Some of the choices made by WRR especially in choosing the appellations seem arbitrary and at times contradictory. It was argued that as long as these choices were made a-priori it is not important that they are not objective. But what can be the explanation of the excellent match in the pair-distance distributions for two experiments based on non-objective a-priori choices?

d. The claim by WRR that the phenomenon is very pointed and that much insight is needed to point down the precise conditions for their phenomenon to occur.

2. One scenario which seems almost impossible by the similarity of the two distributions is that in the first experiment there was some optimization process in setting the ground rules or in choosing the data but not in the second experiment.

3. WRR distinguish between two types of appellations (Rabbi X and others). Within each test the pair-distances coming from Rabbi X appellations (in short X-distances) indeed have different distributions from that of the other pair-distances.

The two distributions of X-distances in the two experiments are very different. [1] The two distributions of the non X-distances from the two experiments are quite similar and when we consider the two distributions of all distances we experience a very close proximity.

This is a little miracle. It fits well, however, with two similar optimizations in which the kernel (see next section) almost coincides with the X-distances for the second experiment and differ from it in the first one.

[1] The phenomenon considered by WRR is much weaker for the X-distances of the second experiment. This agrees with the hypothesis that in the first experiment there was an optimization process in setting the ground rules while in the second experiment most of the optimization was carried out by choosing favorable appellations.

9 The pair-distance distributions for the two experiments separately

Observe the pair-distance histograms which are drawn in [12], p. 437 and [11]p. 4,5 and Figure 2. One striking fact about the pair-distance distributions is that apparently they do not fit at all the suggestion made by WRR that there is a hidden text in the book of Genesis in which we can expect pairs of words which are "related" to be close together. There is a big gap between the claim that something is not random and the claim that there is some hidden order [2] .

The decreasing shape of the histogram even for "bad" distances does not seem to be supported by any reasonable hidden text hypothesis. In particular, what can be the reason for the rare appearances of very large distances (e.g. pair-distances higher than 0.9)? [3] Also, contrary to WRR's rationale to move from P1 to P2 the density of pair-distances below 0.2 is quite constant.

Suppose there was a hidden code. One could expect, for example, that the pair-distances histogram will be a combination of pair-distances for pairs which appear in the hidden text and pair distances for pairs which do not appear in the hidden text. Since three forms of writing each date were chosen (and in several cases two alternative spellings of the same appellations are considered) we can expect a substantial portion of pairs not to be in the hidden text. The distributions we see do not fit at all this possible description.

[2] As parents, we often experience that when we claim to our teen-age child that his (or her) room is in chaos then in response all the mess is shoved to one half of the room. Usually, such a solution is rejected without detailed statistical studies and even before careful peer-review.

[3] This may be described as a hidden "anti-code" - the disappearance of large distances among words with similar meaning. The "bible codes" were extensively used to show that certain events (usually disasters) are encoded in the bible. WRR's work suggests another more peaceful application - to identify disasters that did not and will not happen.

9.1 Comparing with the distribution induced by fixing the P2 score and the number of variables

It seems that the pair-distance histogram exhibits phenomena which are unfavorable to a theory of hidden text but are favorable to the main statistics chosen to verify the research hypothesis. This is not a good sign. It is like somebody who tries to check the hypothesis that a certain university is lowering the academic standards for basketball players. He claims to prove this hypothesis by showing a negative correlation between height and academic achievements. And then it turns out that this negative correlation is supported mainly on short people where there are no basketball players anyway

It is instructive at this point to consider the following distribution: Let $x_1, x_2, \cdots x_n$ be independent random variables uniformly distributed in $[0, 1]$. We are interested in the distribution $\mathcal{M}$ of x_1 conditioned on the property that the P2 statistics of $x_1, .., x_n$ equals B. (In other words we are interested in the distribution of x_1 conditioned on the event that the product $x_1 \cdot x_2 \cdots x_n$ equals a fixed value A.) From equation (1) it is quite easy to compute this distribution precisely. The probability that x_1 is smaller than t is given by

$$(1 - \frac{\log t}{\log A})^{n-1} \qquad ,$$

where A is the product of the x_i's.

Comparing our distribution to the corresponding $\mathcal{M}$-distributions is instructive. (See Figure 3.) The distributions of the experiments are smaller than the corresponding $\mathcal{M}$ distributions in small values and larger on large values. This is opposite to what is expected from a hidden-text assumption. (Like the basketball player example.)

Note that the two $\mathcal{M}$ distributions which correspond to the number of terms and their product in the two experiments are extremely close to each other. (Their sup-norm distance is 0.005.)

Remark: The distributions $\mathcal{M}$ (and some improvements which take into account the actual distribution of "random" pair-distances) seem relevant to what can be expected in other experiments of WRR see [13, 5], where the conjectured optimization process was roughly to choose from a large number of large blocks of pair-distances the blocks with the best P2 scores and to justify this choice later.

9.2 A general model for optimization procedures using the P2 statistics and some simulations

The distributions of pair-distances given by the two WRR experiments does not fit to a hidden code theory. On the other hand, the fact that the two distributions are so close together suggests that there is something robust about these distributions and that perhaps they can be identified. The outcomes of the simulations described below seem promising in this direction.

Consider the following setting for an optimization procedure. The data comes in blocks of sizes between 1 and 6. Each block contains 1-6 independent random numbers uniformly distributed in the unit interval. The probability for a block to have size i is p_i. There are N blocks altogether, M of them (which we call the *kernel*) are fixed and the rest are subject to the optimization process. In the optimization process we simply repeatedly delete the block not in the kernel with the lowest P2 value as long as such a deletion increase the total P2-value of the remaining numbers.

Roughly speaking small block sizes correspond to blunt optimization. The blocks in the kernel just add some uniform distribution.

The principal example we have in mind is pair-distances corresponding to a single appellations. But the model applies to other situations (especially for the first WRR's experiment). For example, for modification or deletions of dates.

The kernel corresponds to pair-distances between appellations which cannot be disputed (and there is no ambiguity in the way they are written) to dates which cannot be disputed. The optimization process apply to the rest of the data.

Of course, this is a very simplified model for the actual optimization process we assume took place. We also simplify greatly the distribution of numbers compared to the original WRR experiments.

Setting and outcomes of the simulations:

1. We fixed the values of p_i according to the number of appellations which contributed i pair-distances, in the two experiments. The corresponding probabilities are $v_1 = (0.15, 0.15, 0.56, 0.09, 0, 0.05)$

and $v_2 = (0.13, 0.3, 0.57, 0, 0, 0)$. We tuned N and M so the median value of the number of remaining numbers and P2 score will be close to the values observed in the experiment.

2. The resulting distributions are described in Figure 4(a) and 4(b). The sup norm distance with WRR's distributions are respectively 0.11 (0.06) and 0.077 (0.29). The values in parenthesis here and below are the probability for a random sample (of the corresponding size) to have a larger sup-norm distance than the sup-norm distance we observe, using the the Kolmogorov-Smirnov approximation. If we *condition* only on distributions with a similar value of P2 score and a similar number of terms the sup-norm distances become 0.09 (0.17) and 0.072 (0.37) respectively.

3. The results are fairly stable if we modify M and N in the neighborhood of the prescribed values. The sup-norm distance between the distribution and the WRR's distribution decreases if the weight of small blocks is increased.

4. We made an experiment to try to evaluate the effect of correlation between the numbers in a given package. For each block we chose at random an interval (mod 1) of length 0.8 and picked at random the numbers of the block in this interval. The value of M dropped substantially. In the second experiment from 168 to 120. The resulting distribution was extremely close to the distributions we witness, the sup-norm distance being 0.05 (0.87) and 0.04 (0.95). (See Figure 4(c).).

9.3 What can be learned from the simulations

The simulations we have made gave us distributions which are fairly close to the two distributions of WRR's experiments respectively.

The simulations also support the qualitative phenomena observed in the WRR data. Large values tend to disappear and the density for small values is fairly constant.

More controlled, systematic and realistic simulations (with a good control of the effect of tuning) and perhaps some analytic study are needed to support and sharpen our explanation to the distributions we see in the two WRR experiment. Such an explanation can be tested on the pair-distance distributions from [9] and [4], and perhaps on other experiments by WRR.

We expect that a simulation which will be based on the actual distributions of pair-distances (rather than uniform distribution) and the actual dependencies of values inside blocks will support and improve the results we obtained.

It is possible to extract the correlation between elements in a given block from the WRR data or even to sample the WRR data in a way in which the marginal distributions are close to uniform. The most realistic simulation will come from computations of the pair-distance function e.g., with respect to random permutations of Rabbis against the original dates as done in [12]. (This is beyond the scope of this paper given our second ground rule.)

At this point our explanation that the pair-distance distributions are the result of an optimization process based on the P2 statistics is the *only* explanation we are aware of to the distributions we observe in WRR's papers. The original hypothesis of WRR is too vague to tell you anything on the distributions and some natural interpretation of it leads to completely different distributions from those seen in WRR's experiments.

10 The work of Bar Hillel, Bar-Natan and McKay

In their experiments WRR had to make a large number of choices. Some of these choices were fixed for the second experiment after being made in the first. Maya Bar Hillel discovered a large number of "degrees of freedom" concerning mainly the dates and various cases where the choices made by WRR were wrong according to their own criteria. However, Witztum, Rips and Rosenberg claim that all their mistakes were innocent and have little effect on the final outcome, and that the choices they made were correct. [4]

Brendan McKay found that WRR did not implement the statistical test agreed upon with Diaconis but another one which increased the significance level by a factor of 100. WRR explain

[4] One of Bar Hillel's finding was the starting point of this work. She observed that in the second list of Rabbis there are a few mistakes according to WRR's own criteria. Two Rabbis should have been omitted and two others should be added (one from the first list). Aumann [2] in a written respond to Bar-Hillel's claims pointed out that when you make these changes the significance level improves by a factor of 40 . There was no reason (even assuming WRR's hypothesis) for such a dramatic improvement so this indicates some instability in the situation. On the other hand I remembered that WRR on other occasions referred to the outcomes of the two experiments as *very similar*.

that this was a misunderstanding and that the statistical test they used in the spirit of Diaconis' suggestion is correct and captures in a better way the phenomenon they were trying to establish.

Dror Bar-Natan discovered that there is a large degree of freedom concerning the choice of the appellations. By searching the "Responsa" database he found that apparently there are more than twice as many appellations for the Rabbis in WRR's second experiment than the number of appellations actually used in the experiment[5](And some of WRR's appellations did not appear at all in the "Responsa" database.) WRR claim that their list of appellations was provided by an independent expert, Prof. Havlin. Prof. Havlin himself explained the criteria for his choices in a document dated October 1996.

BBM further found that the significance level in the WRR paper depends very strongly on the specific choices made by WRR where degrees of freedom exist. This indicates a flawed optimization procedure. Aumann [2] made the point that in many cases this dependence is compatible with the research hypothesis and with WRR claim that their choices are indeed the correct ones.

Another indication of an illegitimate optimization procedure found by BBM is that the excellent significance level reported by WRR strongly depends on several arbitrary choices made by them. This is an indication of an optimization whose objective function depends on these arbitrary choices. (Note: This does *not* indicate that these arbitrary choices themselves were tuned.)

Since the first version of this paper was submitted there were several further developments.

In a letter by Prof. Cohen [8] from Bar Ilan University to Dror Bar-Natan. Prof. Cohen asserts that not only Prof. Havlin's criteria for choosing the appellations are arbitrary and scientifically unsound but also that the specific choices made, violate these criteria. (WRR's preliminary response was that Prof. Cohen is an expert on bible and *not* in the science of bibliography which is more relevant here.)

Bar-Natan and McKay [4] showed also that by choosing appellations for the 32 Rabbis of the

[5] The finding of the vast degrees of freedom concerning the appellations offered a much simpler explanation to WRR experiments than WRR's original explanation and thus it should have been sufficient to nullify their claims. (Especially since these degrees of freedom are not reported in [12].) Shahar Mozes pointed out that the subsequent discussion is similar to the format of debate in the Indian Philosophy where (roughly) in order to refute a certain claim one has to base his argument on the basic assumptions of the claimer [7].

second list in a certain way one can reach the same level of significance reported by WRR in the Hebrew translation of *War and Peace*. BBM claimed that this is possible for quite some time prior to the appearance of [4] and this paper goes much beyond what was promised. In [4] the authors assert that their choices of appellations are in the framework of Havlin's document as much as WRR's original choices are. Moreover, Bar-Natan and McKay further show that several choices of WRR are very unreasonable.

There were several other works trying to make similar experiments to WRR's experiment. Gans [9] checked the proximity of the Rabbis appellations to the cities these Rabbis lived and got significant results. In contrast, Bar-Natan, Gindis, Levitan and McKay see [6], checked the proximity of the names to the *years* of birth and death. They made several hundred experiments and claimed that nothing unusual happened. WRR themselves [13] continued to make several other experiments and claimed to have found further significant evidence for the "hidden code". Bar Natan and McKay [5] study in depth one of their newer experiments.

11 A concluding remark

This work touches only on the tip of the iceberg in the amazing and saddening phenomenon of the "bible code" activity. The phenomenon is not so much about science or religion as it is about unlimited human ambition. I didn't touch at all the question of whether WRR proposed a genuine scientific phenomenon or theory and what should have been the right reaction of the scientific community to start with.

I always regarded the claims that there is some hidden text in the bible as absurd yet harmless. The endorsement by a few mathematicians of Witztum, Rips and Rosenberg's work, the appearance of their paper [12] in a refereed scientific journal, the silence of the mathematical community at large, and Rips' and others use of these "codes" to gain historical and political insights changed matters. My direct involvement was motivated by the recent publications which demonstrated that these codes can be quite harmful. In fact, it is not impossible that the "code" which was claimed to have "predicted" Rabin's assassination has been used for incitement or self-incitement against the late prime minister. The “objective”, “scientific” recognition of this work makes these “bible codes” especially dangerous.

Acknowledgment

It is a pleasure to acknowledge the great help by Danny Braniss, Leo Novik and Michael Ukon in programming. Brendan McKay supplied some raw data which was very helpful. I would like to thank also Maya Bar Hillel, Dror Bar-Natan, Ron Livné, Brendan McKay and Ilya Rips for helpful discussions on the statistical part of this work and the many who helped me greatly in presenting my ideas. Finally, I would like to mention the early work of Nahman Givoli who pioneered the critical study of the ELS activity.

References

[1] The pair-distances from WRR’s paper [11] available at http://sunset.huji.ac.il/ kalai/distances

[2] Y. Aumann, Comments on Maya Bar Hillel’s lecture at Zarka Ma’in. Available at http://sunset.huji.ac.il/ kalai/aumann

[3] M. Bar Hillel, D. Bar-Natan and B. McKay, The Genesis of Equidistant Letter Sequences, in preparation. (This is a paper which is going to contain the authors studies on this matter including the material of [4]).

[4] D. Bar-Natan and B. McKay, Equidistant Letter Sequences in Tolstoy’s ”War and Peace” (draft) available at http://cs.anu.edu.au/ bdm/dilugim/torah.html.

[5] D. Bar-Natan and B. McKay,

[6] D. Bar-Natan, A. Gindis, A. Levitan and B. McKay, Report on new ELS tests of Torah, reply by E. Rips, A rejoinder by the experimenters. all can be found in http://cs.anu.edu.au/ bdm/dilugim/torah.html

[7] S Biderman, Indian Philosophy (in Hebrew), Ministry of defense publishing house, Tel Aviv, 1980.

[8] A letter from Prof. M. Cohen, available at http://cs.anu.edu.au/ bdm/dilugim/torah.html.

[9] H. Gans,

[10] S. Sternberg, Comments on the *Bible Code*, Notices of the AMS, September 1997.

[11] D. Witztum, E. Rips and Y. Rosenberg, On Equidistant Letter Sequences in the Book of Genesis, preprint, 1987.

[12] D. Witztum, E. Rips and Y. Rosenberg, On Equidistant Letter Sequences in the Book of Genesis, Statistical Science, 9 (1994), 429-438.

[13] D. Witztum, E. Rips and Y. Rosenberg, On Equidistant Letter Sequences in the Book of Genesis II, preprint (1997), 429-438.

[14] S. Wilks, *Mathematical Statistics*, Wiley, New-York, 1962.