The Decline of Violent Conflicts: What Do The Data Really Say?

Pasquale Cirillo and Nassim Nicholas Taleb[1]

Nobel Foundation Symposium 161: The Causes of Peace

Summary: We propose a methodology to look at violence in particular, and other aspects of quantitative historiography in general, in a way compatible with statistical inference, which needs to accommodate the fat-tailedness of the data and the unreliability of the reports of conflicts. We investigate the theses of "long peace" and drop in violence and find that these are statistically invalid and resulting from flawed and naive methodologies, incompatible with fat tails and non-robust to minor changes in data formatting and methodologies. There is no statistical basis to claim that "times are different" owing to the long inter-arrival times between conflicts; there is no basis to discuss any "trend", and no scientific basis for narratives about change in risk. We describe naive empiricism under fat tails. We also establish that violence has a "true mean" that is underestimated in the track record. This is a historiographical adaptation of the results in Cirillo and Taleb (2016).

Preamble

The first theory of "long peace" is as follows. In 1858, one H.T. Buckle wrote:

> That this barbarous pursuit is, in the progress of society, steadily declining, must be evident, even to the most hasty reader of European history. If we compare one country with another, we shall find that for a very long period wars have been becoming less frequent; and now so clearly is the movement marked, that, until the late commencement of hostilities, we had remained at peace for nearly forty years: a circumstance unparalleled (...) The question arises, as to what share our moral feelings have had in bringing about this great improvement. (Buckle, 1858) .

[1] Delft Institute of Technology and Tandon School of Engineering, New York University. The contributions of the authors to this paper and the general statistical studies associated with it, are equal. We thank Captain Marc Weisenborn for his diligent and indefatigable work in collecting, checking, and comparing data.

Buckle was perhaps "right" (with minor hick-ups) for another five decades, but moral feelings or not, the century following Mr. Buckle's prose turned out to be the most murderous in human history.

The first – obvious – problem is that Buckle made a severely flawed risk assessment. The second is that he felt obligated to mount a narrative entailing "moral feelings" for what he perceived were the changes in the environment. Note that Buckle was placing himself in the tradition of "scientific" social science of Auguste Comte.

Another event. In 2004, Ben Bernanke, then a member of the board of the Federal Reserve Bank of the United States proclaimed that economic life was undergoing a "great moderation", on the basis of an unprecedented stability of economic variables (Taleb, 2007, 2010). Like Buckle, he found the reasons for that. The theory became the norm until the crisis of 2007 when we experienced a similar revision of belief.

***

This article is organized as follows. First we present the problems associated with historical analyses of violence. Second, we discuss the quantitative approaches since Richardson (1948) and present the statistical flaws and methodological errors in the widely held theories such as those in Pinker (2011). Third, we discuss our approach. Fourth, we provide a slightly more technical backup. Fifth, we give our verdict.

Some Problems in Quantitative Historiography

Studying the history of violence to detect trends and changes over a time period is a non-trivial task for a scientist constrained by rigor. We list five problems that are particular to violence, but may be universal to any form of quantitative and statistical historiography.

Problem 1: Fat tails. First, we are dealing with a "fat-tailed" phenomenon. We define violence seen quantitatively as either "fatalities over a specific time period" or "fatalities per specific event" and both are fat tailed variables. What characterizes fat tailed variables? These have their properties (such as the average) dominated by extreme events, those "in the tails". The most popularly known version is the "Pareto 80/20" (80 % of the people in Italy, Pareto noticed, own 20% of the land, and vice versa, which by recursion leads to 1% of the people owning 53% of the land).

These tools are not just "changing the color of the dress", but they require a new statistical framework and a different way of thinking, going from the tail to the body (standing the usual statistical logic, which consists of going from the body to the tail, on its head) –and the great majority of researchers who are trained in statistics are not familiar with the branches of the discipline and theorems needed for fat tails (see Taleb 2007/2010, 2016). To add to the problem, our examination shows that war turned out to be the mother of fat tails, far worse than the popular 80/20 rule: there are few phenomena such as fluid turbulence or thermal spikes on the surface of the sun that can rival the fat-tailedness of violence. Further, historical data are temporal (spread out over time) and statistical analyses of time series (such as financial data) require far more sophistication than simple statistical tests found in empirical scientific papers. For instance one cannot blindly use the same methods to compute the statistical properties of city size and the time series of war –since in the latter case the observed properties depend on our survival (say a 1960 nuclear war would have prevented us from having this discussion), hence restriction apply to what can or cannot be

inferred (there is a difference between ensemble probability and time probability, though not always, and the effect of the bias needs to be established).

Problem 2: Boundedness. Not only are we dealing with extreme fat tails, but the effect is bounded quantitatively, with an almost precisely known upper limit –no war can kill more than the population of the planet. This brings in an additional mathematical complication, since all techniques for fat tails requires an infinite support for the variable. The boundedness requires some formulaic adjutment to the statistics of violence –which, as we show in some detail, has had so far a mathematically and statistically naive literature. But it is not all bad news since there will be a statistical "mean", which, carefully interpreted, can help in the analysis – while naive statistical analysis produces an "infinite" or "undefined" mean.

Problem 3: Reliability of historical data. The analysis needs to incorporate the unreliability of historical data –there is no way to go back and fact-check the casualties in the Peloponesian war and we rarely only have more than one side to the story. Estimates of war casualties are often anecdotal, spreading via citations, and based on vague computations, almost impossible to verify using period sources. Even more recent events, such as the Algerian war of the 1960s have two polarized sides to the story; and, in some cases such as the Armenian-Syriac genocide of 1915-1917, the numbers do not converge, and have actually been diverging over a century. This unreliability requires taking into account another layer of uncertainty. In addition, the analysis must consider that many wars went unreported –we just do not know how many, and such a number is itself a random variable. There exists even a third layer of uncertainty: the number of gaps between wars can be treated as a random variable, and its effect must be taken into consideration in the interpretation of the results.

Problem 4: The definition of an event. A hurdle can be the precise definition of an event, which appears to be a function of the sophistication of the historian and his or her closeness to one side involved in it. We mentioned earlier the underlying statistical variable defined as "fatalities per specific event", but the very definition of an event matters –and the analysis should not be fragile to the specification. For instance, Pinker (2011) treats as a single event the "Mongolian invasions" which lasted more than a century and a quarter. This swelled the numbers *per event* over the Middle Ages and contributed to the illusion that violence has dropped since, given that subsequent "events" had shorter durations. Effectively the main sources such as Philips and Axelrod's 3-volume *Encyclopedia of War* list numerous conflicts in place of "Mongol Invasions" –the more sophisticated the historians in a given area, the more likely they are to break conflicts into different "named" events and, depending on historians, Mongolian wars range between 12 and 55 conflicts. If some Mongolian-centric Pinker's counterpart wrote about European wars, he would have bundled the period from the Franco-Prussian war to WW II as "Northern European or Western wars".

Not only are "Named" conflicts – part of what Richardson called "deadly quarrels" –an arbitrary designation that, often, does not make sense statistically, but a conflict can have two or more names; two or more conflicts can have the same name, and we found no satisfactory clear-cut hierarchy between war and conflict. Our solution is 1) to treat events as the shorter of event or its disaggregation into units with a maximum duration of 25 years each, which corresponds to a generation, and 2) perform a study of statistical robustness (on which much, further down) to assess whether other time windows and geographic redefinitions produce different results.

Likewise, the data makes it hard to assess whether the numbers include people who died of side effects of wars –say for example it makes a difference whether the victims of famine from the siege of Jerusalem are included or not in the historical figures.

Problem 5: Units for the Analysis. Should one consider, in a trend analysis, raw or relative numbers, that is actual number of people killed or their ratio to the total population? Given that the population of the earth has been increasing over time, a constant rate of violence would give the illusion of rise in casualties.

Our paper Cirillo and Taleb (2016) set to construct the most statistically robust picture of historical violence for "named conflicts" we could make *under the constraints* given by these five problems. To deal with the first problem (Fat tails) we made use of a branch called Extreme Value Theory. For problem 2 we had to develop a technique to allow boundedness in the analysis and publish as standalone as it applies to other similar problems (Taleb, 2016, Taleb and Cirillo, 2015). To deal with problems 3 through 5 we relied on various methods from the branch of robust statistics.

Fat Tails and Theories of Violence

The first main attempt to model violence using power laws was done by the polymath Lewis Fry Richardson, in 1948, when he fit a visual power law using Zipf Plots to data between 1820 and 1945. Zipf plots are a visual technique named after George Zipf who popularized Pareto's law by applying it to phenomena such as linguistics (word frequency) , demographics (size of cities), and others. Richardson himself came from discoveries of self-similarity and scaling in nature, particularly coastlines and some turbulent phenomena (Mandelbrot 1983).

Intuitively, we can explain power laws as follows: $\frac{\#events\ killing\ more\ than\ 100K}{\#events\ killing\ more\ than\ 50K}$ approximately equal to $\frac{\#events\ killing\ more\ than\ 200K}{\#events\ killing\ more\ than\ 100K}$. This "scalability" is crucial as it makes the law both more intuitive and tractable. More precisely there is an "exponent" in the tails, such that, for a large deviation K,

$$Probability\ of\ exceeding\ K = proportional\ to\ K^{-\alpha}$$

where "alpha" is the tail exponent and the major determinant to the shape of the distribution. The lower the alpha, the fatter the tails. Note here a helpful property that the alpha does not change if one takes half the data and doubles it. It means the alpha is *robust to many mistakes in data*. Also note that the proportionality holds in the tails i.e. K very large, but not necessarily for smaller values.

Many research papers subsequently confirmed the power law or "Paretianity" of the data (although Cirillo 2013 shows than many phenomena identified as power laws are in fact lognormal, though still fat-tailed because of their high variance) –but the practice in that field was to find a mechanism that justifies a statistical law prior to adopting it. Mechanisms abound, and after the works of Mandelbrot (1983) linking the phenomenon to fractal geometry (as Richardson did), many branches of "complexity theory" were born. Cederman (2003) looked at a broader set and "justified" the process thanks to a class of models called *agent-based* that have been known to produce power laws. Cellular automata and models of interaction between agents have now flourished providing us a large computational-based

modeling apparatus; it suffices to specify the conditions and the mathematical and computational framework in place allows us to check whether power laws emerge or not (see Wolfram, 2002 and Mathematica's programming language; note that these models are not analytic-functional but algorithmic requiring computational methods). It is not part of this analysis to discuss these models –our domain is statistical inference not model building; we focus on the statistical backup that allows the rejection or "acceptance" of some models.

We note that all "fat tails" are not power laws, there are distributions that produce concentration –it is just that power laws are more natural because of their scalability and appeal to users as they are very tractable analytically. But there are claims one should never casually make in the presence of fat tails, power laws or not. Let us now put together the problems of statistical inference under fat tails.

Fat Tails , Long Peace, and the Foundational Principles of Statistical Inference

Pinker (2011, 2011b) started the promotion of an idea that violence has dropped, with similar to Buckle –eerily similar— an invocation to the various moral values causing what he calls "the obsolescence of major wars" (Pinker 2011b). He writes, Buckle-like, "The most promising explanation, I believe, is that the components of the human mind that inhibit violence — what Abraham Lincoln called "the better angels of our nature" — have become increasingly engaged."

It is important to discuss that book because it has been cited as "evidence" for the drop in violence across political science. Pinker deals with the phenomenon of violence, and its manifestations at different scales, from homicides and rapes, to riots and wars, from death penalty and torture issues, to civil rights violations and denial. To explain and sustain his vision about the general decline in violence, Pinker develops the metaphor of a constant battle, within humanity, among some ``Inner Demons", like for example revenge, sadism and ideology, and some ``Better Angels", such as empathy, self-control and reason (even if it is not completely clear, at least to us, what it is that Pinker calls reason). Demons are mainly expressions of atavic feelings and compulsions, which are related to the original beast in us. Angels are a result of civil evolution and reason development. And since civilization seems to be an unstoppable process, Angels are bound to win the battle.

Using a sort of meta-analysis, relying on others' results, Pinker collects a bunch of figures to support the idea of a decline in violence in the history of humanity.

To armed conflicts, in all their possible expressions, he devotes two chapters: the fifth, ``The Long Peace", and the sixth, ``The New Peace". In the specific case of wars, he relies on previous analyses by, for example, Cederman (2003) and Richardson (1960). But the way in which he reads and interprets the results of scholars like Richardson reveals an attempt of bending empirical evidence to his own theory, e.g. when he deals with the Poisson nature of the *number* of armed conflicts over time. As we also find out in our data analysis, consistent with Richardson (1960), there is no sufficient evidence to reject the null hypothesis of a homogenous Poisson process, *which denies the presence of any trend in the belligerence of humanity.* Nevertheless, Pinker refers to some yet-unspecified mathematical model that could also support such a decline in violence, what he calls a "nonstationary" process, even if data look the way they look. It is on the basis of this and other apodictic statements that Pinker builds his narrative about violence.

Pinker, in addition builds a theory that is not at all statistically robust to problems 3 and 4 of the previous section–such as as changes in the An Lushan estimates or the granularity of the Mongolian named conflict.

But at the core, Pinker's severe mistake is one of standard naive empiricism –basically mistaking data (actually absence of data) for evidence and building his theory of *why* violence has dropped without even ascertainting *whether violence did indeed drop*. This is not to say that Pinker's socio-psychological theories can't be right: they are just not sufficiently connected to data to start remotely looking like science. Fundamentally, statistics is about ensuring people do not build scientific theories from hot air, that is without significant departure from random. Otherwise, it is patently "fooled by randomness". And we have a very clear idea what departure from random means.

For fat tailed variables, the conventional mechanism of the law of large numbers (on which statistical inference reposes) is considerably slower and significance requires more data and longer periods. Taleb (2016) shows that the Pareto 80/20 takes 10^{13} more data than a corresponding Normal distribution that is ubiquitous in textbooks if one looks at the sample average. Simply, the sample average is not a good estimator of the "true" mean; it has what is called a small sample bias when data is one-tailed (i.e. can only take either positive or negative values, as is the case with violence). In other words, not only do we need a lot of data to know what's going on, but , as in the case of violence, we should expect that the mean violence as measured in sample to be lower than the true mean. The statistician would never measure the mean in-sample but use and we will resort to "back-door" methods and more rigorous maximum-likelihood techniques.

Ironically, there are claims that can be done on little data: inference is asymmetric under fat-tailed domains. *We require more data to assert that there are no black swans than to assert that there are black swans* hence we would need much more data to claim a drop in violence than to claim a rise in it .

Finally, statements that are not deemed statistically significant –and shown to be so –should never be used to construct scientific theories. Descriptive statistics, though deemed unscientific and anecdotal, can be useful for exploratory discussions, but not with fat tailed processes when the random variable entails exposures rather than binary outcomes.

These foundational principles are often missed because, typically, social scientists' statistical training is limited to mechanistic tools from thin tailed domains. In physics, one can often claim evidence from small data sets, bypassing standard statistical methodologies, simply because the variance for these variables is low (or the process has a strong theory verified on a high signal to noise ratio). The higher the variance, the more data one needs to make statistical claims. For fat-tails, the variance is typically high and underestimated in past data. And, as we showed it drops very slowly under averaging (the law of large numbers means the sample average becomes less and less volatile as one increases data, typically at the rate of the square root of additional data counts; this is not the case here).

The second –more serious –error Pinker made in his conclusion is to believe that tail events and the mean are somehow different animals, not realizing that the mean *includes* these tail events. *Further, for fat-tailed variables, the mean is almost entirely determined by extremes. If you are uncertain about the tails, then you are uncertain about the mean*. It is thus incoherent to say that violence has dropped but maybe not the risk of tail events; it would be like saying that someone is "extremely virtuous except during the school shooting episode when he

killed 30 students", or that nuclear weapons are very safe as they only kill a small percentage of the time.

Our methodology

The Data: We selected all available observations over the period 1-2015 AD, usually armed conflicts with more than 3,000 casualties in absolute terms, counting both soldiers and civilians.

There are a few exceptions in our data set, some events that cannot be considered standard armed conflicts in the definition of Wallensteen and Sollenberg (2001). They are some of the bloodiest dictatorships of history, such as Stalin's regime. This choice has been made to be consistent with other works about war victims and violence (e.g. Mueller, 1989; Pinker, 2011).

The data come from different sources, such as Phillips and Axelrod (2004) *Encyclopedia of Wars,* and Necrometrics (2015), with some considerations of selections by Berlinski (2009), Goldstein (2011), Mueller (1989), Pinker (2011), and White (2013). For some online resources like Necrometrics (2015), we double-checked the data against the cited references.

The first observation in our collection is the Boudicca's Revolt of 60-61 AD, while the last one is the still-open international armed conflict against the Islamic State of Iraq and the Levant. We ended concentrating on 565 pieces.

A natural question is why we have chosen to impose a 3000-casualties threshold, when collecting the observations about armed conflicts. We have three main reasons:

- We do not need smaller casualties to get the properties, as smaller casualties do not affect the average (Richardson himself noted, "Anyone who tries to make a list of "all the wars" encounters the difficulty that there are so many small incidents, that some rule has to be made to exclude them." (Richardson, 1948). The higher the threshold, the fewer the observations and the lower the noise and imprecision.
- The main object of our concern is tail risk, that's the risk of major destructive conflicts. The statistical techniques we use to study this type of extreme events require the imposition of thresholds for all the approximations to hold.
- Conflicts with many victims are more likely to be registered and studied by historians. It would be impossible to have reliable information about "small" battles with tens of victims. Empirically, 3000 victims proved to be a good selection threshold for other aspects of the analysis.
- A 3000-casualties threshold gives us a better confidence about the estimated number of casualties, thanks to the possibly larger number of sources to compare. However, the risk of over-exaggeration, especially for the large conflicts of antiquity is something we had to take into consideration.

In order to be consistent with the sociological literature on armed conflicts, we have used different types of data in our analyses:

- *Actual data*, i.e. data as collected by historians. Statistically speaking these are the raw data.
- *Rescaled data*, i.e. casualties expressed in terms of today's world population, in order to have comparability in terms of relative impact of wars.
  Rescaled data are obtained by dividing the number of casualties in a given year by

the world population in that year, and then multiplying everything by today's world population.
When rescaling data, we have used the population estimates of Klein and van Drechts (2006), and United Nations (1999, 2015).

- *Transformed-rescaled data*, that is data obtained via the so-called dual transformation, whose aim is to deal with the boundess/unboundeness of the support of the distribution of war casualties (Taleb, 2016; Taleb and Cirillo, 2015). This transformation is more technical, but, as we briefly explain in the section about Methods, it is meant to correct for an interpretation error about the apparent infinite-mean nature of war casualties' data.

Interestingly, our results hold notwithstanding the definition of data. Numerical estimates may vary, but the qualitative interpretation stays the same.

Data Problems: First of all, it is important to notice that our data, especially for what concerns antiquity, are likely to suffer from selection/historiographical bias. It was in fact not possible to collect observations about the conflicts taking place in the Americas and Australia, before their "discovery" by European conquerors. Naturally, this lack of evidence does not mean that nothing happened in those areas in the past.

Similarly, because of problems with sources, we probably miss some conflicts of antiquity in Europe, or, say, in China in the sixteenth century. However, we can assume that the majority of these conflicts are not in the very tail of the distribution of casualties, say in the top 10 or 20%. It is in fact not really plausible that historians have not reported a conflict of 1 million casualties (or more), so that such an event is not present in our sources.

Dealing with historical data, some dating back to the first century AD, also requires some attention, because of the probably problems of inconsistency and lack of uniformity in the attribution of casualties by historians. It can be difficult – if not impossible – to distinguish casualties from direct violence from those arising from such side effects as contagious diseases and hunger.

We mentioned earlier that reports were highly source dependent with an impossibility to fact-check. Some data, such as the An Lushan rebellion, estimated by Pinker (2011) to have killed 36 million people (around 430 million by today's population) are highly dubious (Durant, 1960) –and help perpetuate the impression that the world is "less violent". It may have been 13 million, and the numbers were the result of the census and dispersion of officials in revenue department. (Fitzgerald, 1935, BBC 2012)

Data aggregation is another issue. We said that conflicts such as the so-called "Mongolian Invasions" are nothing more than artificial designations, which need to be treated carefully, as synthetic observations. These events are in fact artificial containers created by historians to aggregate those battles sharing important historical, geographical and political characteristics, but that never really existed as a single event. For historical/historiographical reasons, these events tend to be more present in antiquity and the Middle Ages, thus possibly causing a naive overestimation of the severity of conflicts in the past. Even among these aggregations there are major differences: WW1 and WW2 naturally also involved several tens of battles in very different locations, but these battles took place in a much shorter time period, with no major time separation among them.

Curing the data problems: All these data problems can be dealt with by considering each single observation in our collection as an imprecise estimate, in the definition of Viertl (1995). Our technique for robustness is as follows. Using Monte Carlo methods , and assuming that the real number of casualties in a conflict is uniformly distributed between the minimum and the maximum estimate in the available historical records, the tail exponent ξ, the quantity that governs the tail of our war casualties' distribution, is not affected, apart from the negligible differences in the smaller decimals. We did another battery of tests for other variables. (See Figures 1 and 2 for an idea of how we conducted the tests.)

From a statistical point of view, the methods we use to study tail risk are robust. In simple words, this means that our results – and the relative interpretations – are immune to small changes in the data. Even more: our results cannot be reversed on the basis of a few observations, added, removed or "corrected". A thorough analysis of robustness shows that our estimates are preserved and would replicate even if we missed one third of the data.

To conclude this section about "data problems", we think it is important to stress that our data set, despite its evident temporal connotation, does not form a proper time series. It is in fact trivial to notice that the different conflicts of humanity do not share the same set of causes. Battles belonging to different centuries and continents are not only independent, but also surely have different origins. In statistical words, we cannot assume the existence of a unique *conflict generator process*, as if conflicts were coming from the same source.

For this reason, we believe that performing time series analysis on this kind of data is useless, if not dangerous, given that one could extrapolate misleading trends, as done for example in Pinker (2011). How could the An Lushan rebellion in China (755 AD) be dependent on the Siege of Constantinople by the Arabs (717 AD), or have an impact on the Viking Raids in Ireland (from 795 AD on)?

Notice that we are not saying that all conflicts are independent: during WW2, the attack on Pearl Harbor and the Battle of France were not independent, notwithstanding the spatio-temporal divide, and that's why historians merge them into one single event, as we have already noticed before. And while we can accept that, historically, most of the causes of WW2 are related to WW1, it is better to avoid translating this dependence when studying the number of casualties: it would be quite absurd to believe that the number of victims in 1944 had anything to so with the death toll in 1917. How could the magnitude of WW2 depend on WW1?

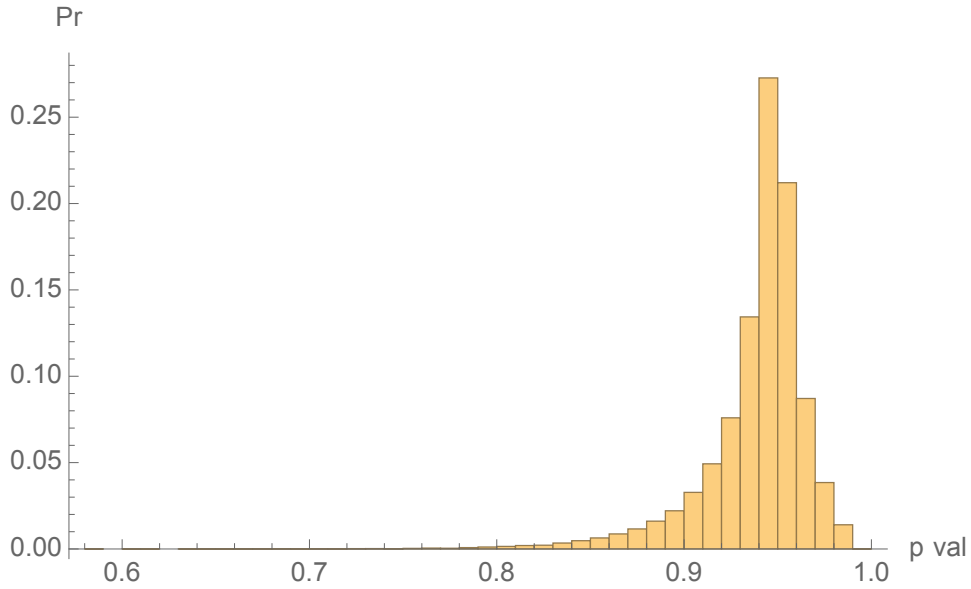

Figure 1 How we tested our robustness to the reliability of historical reports. We create 100,000 different histories as uniform random numbers between high and low estimates from the data sets (which under aggregation appear Gaussian) and check if re-combinations leads to different results (we used the p-value, actually 1-p-value for the scale parameter not because we rely on p-values but because p-values are extremely sensitive to changes in data). Some histories include Pinker's exaggerated numbers for the Ann Lushan rebellion, others don't.

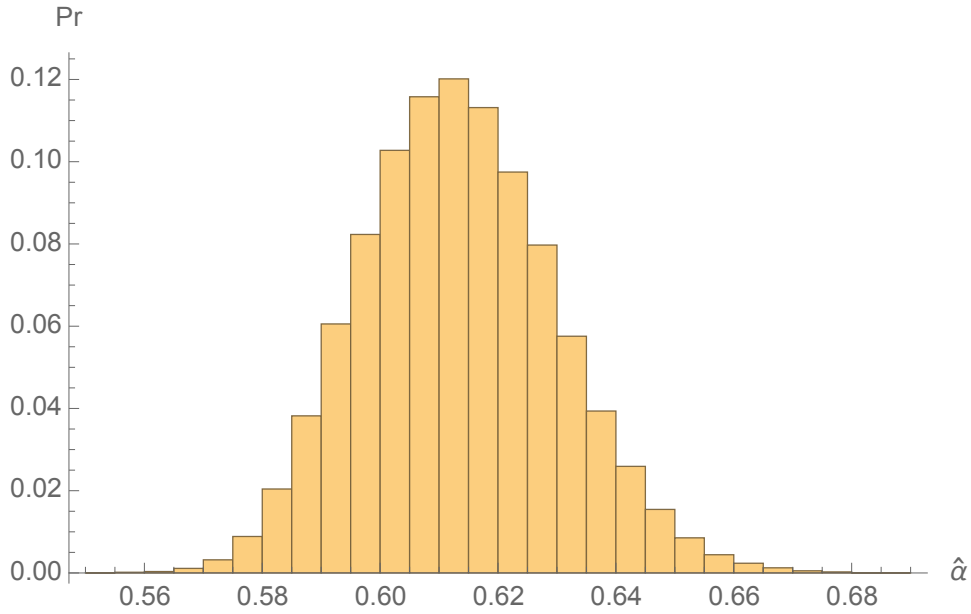

Figure 2 The tail exponent from maximum likelihood, not EVT, is invariant to errors in the reporting of conflicts.

Methods

Since we are interested in estimating the tail risk of violent conflict, that is the risk of large destructive wars and armed conflicts, we use tools from extreme value theory (EVT) to

understand the behavior of the right tail of the distribution of war casualties. EVT is a branch of statistics dealing with extreme and rare events, in the form of maxima and minima (Gumbel, 1938; Embrechts et al., 2003; de Haan and Ferreira, 2006).

Within the broad field of EVT, we mainly use the so-called Generalized Pareto (GP) approximation (Balkema and de Haan, 1974; Pickands, 1975), according to which all the exceedances above a high threshold, if this threshold is correctly chosen, tend to follow a GP distribution, a skewed distribution, which can be characterized by fat tails.

The function form of a GP distribution is

$$GPD(z;\xi,\sigma,u) = \begin{cases} 1-\left(1+\xi\frac{z-u}{\sigma}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1-e^{-\frac{z-u}{\sigma}} & \xi = 0 \end{cases},$$

where $z \geq u$ for $\xi \geq 0$, and $u \leq z \leq u - \sigma/\xi$ for $\xi < 0$, with $u, \xi \in \mathbb{R}$ and $\sigma > 0$.

The parameter ξ is the most important one for us, as it controls for the fatness of the right tail. The larger ξ, the fatter the tail. For $\xi > 0.5$, the GP distribution has an infinite variance. For $\xi > 1$, even the mean is not finite.

For what concerns war casualties, we find that the Paretian tail is actually so fat that, from a theoretical point of view, the mean of the distribution is not finite ($\xi > 1$). In simple words, this means that the tail risk is so large that one single event, one single war, could destroy the whole humanity (7.3 billion people).

In reality, things are a little bit more complicated, because data can be misleading, even when approaching a problem using the correct methodology (EVT, when studying tails). While we show that the tail risk of violent conflict is actually large – much larger than what one could simply infer using standard descriptive and inferential statistics (not appropriate in this case), we also point out that it cannot be infinite, as data tend to suggest naively using EVT. In fact, no single conflict can kill more than the whole world population. This implies the presence of an upper bound that we can use to correct our estimates. The *dual distribution* approach, based on a special log-transformation of data, is the way in which we deal with apparently infinite mean phenomena like war casualties. For more details about our methodology, especially for the use of the dual distribution, we refer to Cirillo and Taleb (2016).

Once again, it is worth underlining that all the statistical methods we use are robust, that is to say they tend to be immune to even to non-trivial changes in the data (a third of reported events in our data could change significantly, and still our results are preserved).

The distribution of war casualties: basic facts

When looking at the distribution of war casualties, the first thing we notice is that it is highly skewed with a very fat right tail. Figure 3 contains a simple histogram using actual data:

while most armed conflict generate a few thousands victims[2], on the left-hand side of the picture, a few conflicts cause millions of casualties, with WW2 totaling between 48.5 and 85 million victims, depending on the source (in the graph we show the average: 73 mio).

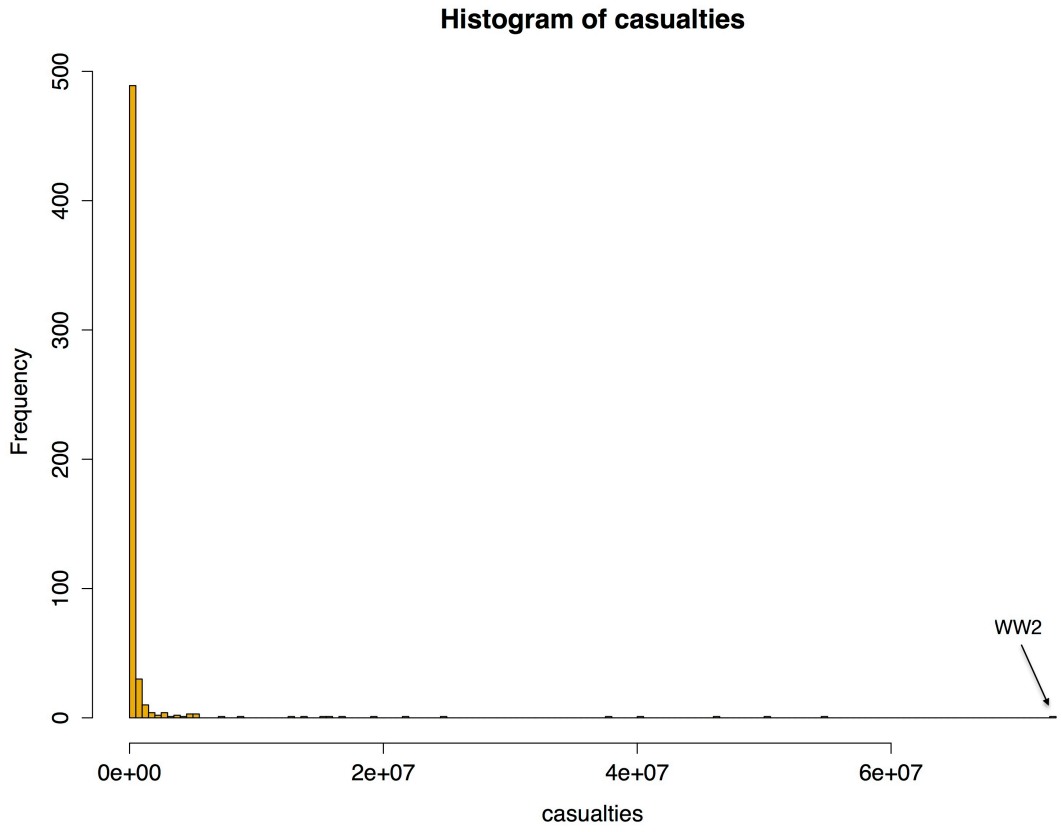

Figure 3: Histogram of war casualties using raw data.

The average number of casualties in our sample is 1,067,568. The in-sample standard deviation is 5,738,541. However, since the standard deviation is not a reliable measure under fat tails, given that the theoretical variance may not exist (Embrecht et al., 2003), we also provide the mean absolute deviation, or MAD[3]: 1,747,869. This number shows the extreme volatility of war casualties, something compatible with a fat-tailed phenomenon.

[2] Please notice that we are not giving any ethical judgement. When we say « a few victims », we think of them as statistical data, numbers. From an ethical point of view, one victim is already too much.

[3] The MAD is the mean of all the absolute deviations of each data point from the sample mean. Formally: $MAD(x)=E|X-E[X]|$.

Other useful statistics are the median (40,000), the first quartile (13,538) and the third quartile (182,000): 75% of all armed conflicts generate less then 182,000 casualties, and still the sample mean is almost 6 times larger!

Figures 4 and 5 show the number, and the relative magnitude with respect to world population, of war casualties over time, when using two different definitions of data: actual and rescaled observations.

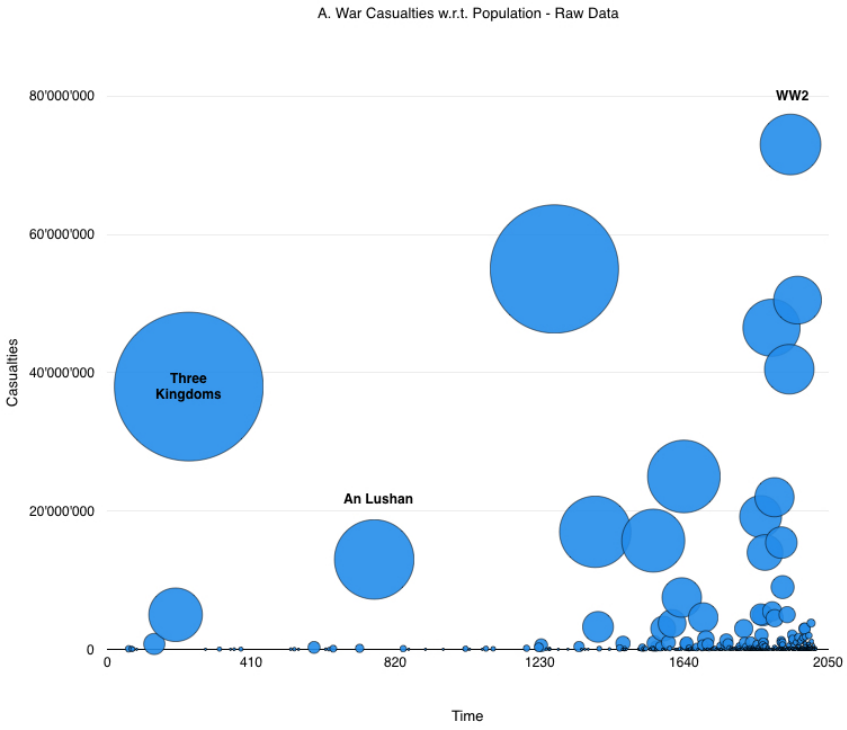

Figure 4: War casualties over time w.r.t. world population using actual data.

It is interesting to notice how the choice of the type of data may lead to different interpretation of trends and patterns. From rescaled data (Figure 3), one could for example superficially infer a decrease in the number of casualties over time. It is in fact obvious that rescaling data will tend to inflate past observations, as already noticed by Epstein (2011), who also object that rescaling may generate paradoxical situations. Citing him: "[...] why should we be content with only a relative decrease? By this logic, when we reach a world population of nine billion in 2050, Pinker will conceivably be satisfied if a mere two million people are killed in war that year".

As to the number of armed conflicts, both figure 5 and 7 seem to suggest an increase of belligerence over time, since most events are concentrated in the last 500 years or so. This is very likely just an illusion, probably due to a reporting bias for the conflicts of antiquity and early Middle Ages. It is certainly easier to obtain decent information about more recent

conflicts, that is why we have many rather precise observations in the last decades and centuries, with respect to what happened in the third century AD.

To correctly study the tail risk of armed conflicts, we need to understand whether our data really exhibit a Paretian right tail, as suggested by Figure 4. The answer is affirmative. For example, in Figure 6, we see that, on a log-log scale, the distribution of war casualties shows a clear linear behavior in the right tail. Figure 6 is a Zipf plot (mentioned earlier), and it represents a heuristic tool to look for Paretianity in data (Cirillo, 2013), by using a property of the survival function of a Pareto distributed random variable. Many other plots can be used to verify the presence of a fat right tail, such as Maximum-to-Sum and mean excess plots, and we refer to Cirillo and Taleb (2016) for more details and discussions.

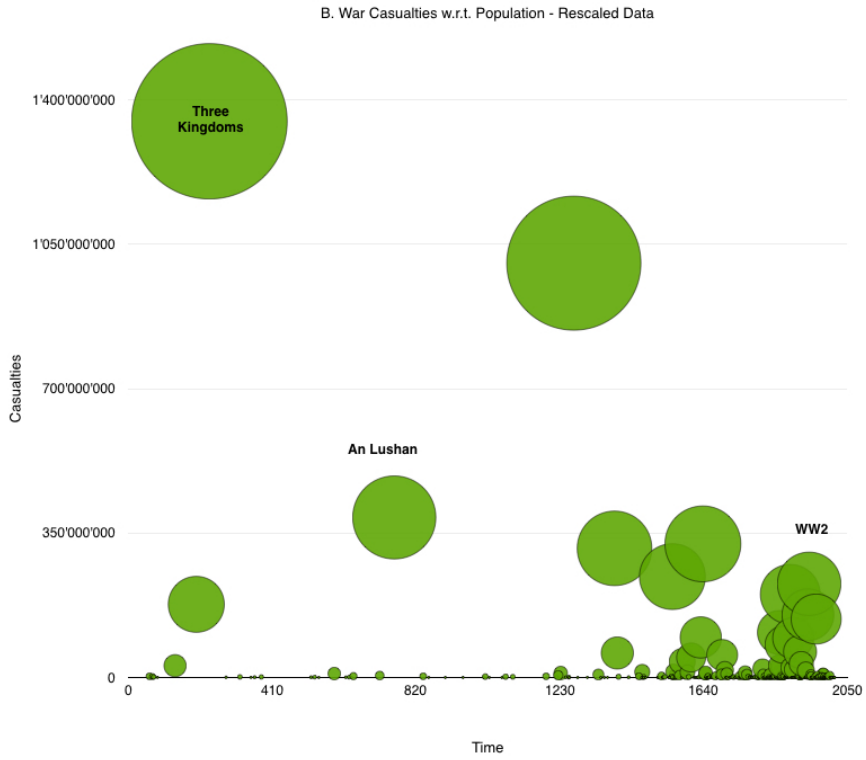

Figure 5: War casualties over time w.r.t. world population using rescaled data.

Figure 6 not only confirms the idea of a Paretian right tail, but it also suggests that the whole distribution of war casualties may be in the domain of attraction of a fat-tailed distribution. The linearity of the survival function starts indeed from the very left-hand side of the plot. In addition Figure 6 shows how the right tail tends to close down a little bit. This is a phenomenon commonly known as finite sample bias. It is in fact highly improbable to be able to observe a sufficient number of maxima in the data, so that the tail decreases linearly until the end.

Table 1 contains some information about the occurrence of armed conflicts over time. For example, if we look at events generating at least 500,000 victims, we discover that, when

using raw data, we have to wait an average of 24 years to observe such events. The corresponding mean absolute deviation is 33 years. If, on the contrary, we use rescaled data, the average inter-arrival time is 10 years, with a MAD of 12.

Threshold	Avg Raw	MAD Raw	Avg Rescaled	MAD Rescaled
500k	24	33	10	12
1 mil	34	48	13	16
2 mil	57	73	20	24
5 mil	93	117	34	43
10 mil	136	139	52	61
20 mil	252	267	73	86
50 mil	372	362	104	114

Table 1: Average inter-arrival times and their mean absolute deviation, in integer years, for different casualties' thresholds.

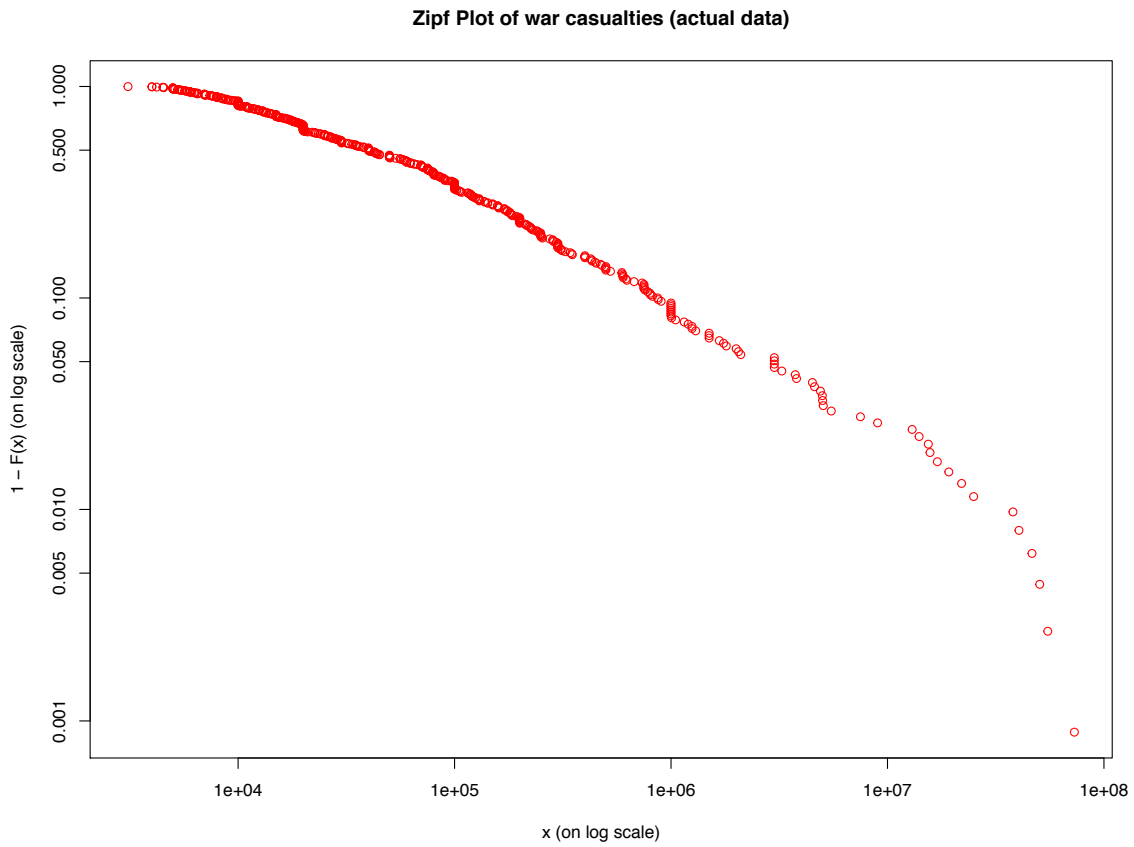

Figure 6: Zipf Plot (log-log plot of the survival function) of war casualties using actual data.

If we increase the threshold[4], and consider conflicts with at least 5 million casualties, we need to wait 93 or 34 years, depending on the data definition. Intuitively, the bloodier the conflict, the longer the inter-arrival time. For a conflict with at least 50 million victims, a very extreme and hopefully rare event, the average inter-arrival time is 372 years, using raw data, with a MAD of 362.

All this tells us that the absence of a conflict generating more than – say – 20 million casualties in the last 20 years is highly insufficient to state that its occurrence probability has decreased over time, given that the average inter-arrival time is 252 years (73 for rescaled), with a MAD of 267 (86 for rescaled) years! Unfortunately, we still have to wait quite some time to say that we are living in a more peaceful era; the actual data we have are not in favor nor against a structural change in violence, when we deal with war casualties. Very simply: we cannot say.

[4] The thresholds in Table 1 are just arbitrary, and meant to give useful information in a compact table. Other thresholds can be chosen, but one general monotone behavior can be observed, in accordance with intuition: the higher the threshold, the longer the average inter-arrival time.

The tail risk of armed conflicts

As said before, using extreme value theory, and in particular the generalized Pareto approximation, we can extrapolate information about the tail risk of armed conflicts.

The estimation of the parameters of a GPD can be performed in different ways, and we refer to Embrechts et al. (2003), or de Haan and Ferreira (2006) for more details. With our data, the best results are obtained using maximum likelihood (Cirillo and Taleb, 2016).

Table 2 contains our estimates for the generalized Pareto approximation for the distribution of war casualties, using both actual and rescaled data. For both definitions, we can see that the thresholds above which the GPD approximation holds are definitely larger than the original 3000 casualties we have imposed in collecting the data. However, in both cases, almost 60% of all the observation lie above the two thresholds. Since we are interested in the risk of very large conflicts, dealing with the top 60% of all conflicts is definitely more than sufficient for our purposes[5].

The most interesting result of Table 2 is the qualitative information we can extrapolate from ξ (decimals are much less relevant). For both actual and rescaled data, we clearly see that $\xi > 1$. This indicates the presence of an infinite mean phenomenon, that is a phenomenon whose realizations can be so large and erratic that the mean is not a reliable quantity. And if the mean of the GPD is not finite, then the mean of the whole distribution of war casualties must be infinite. In fact, since the tail mean is a component of the whole distribution mean, if the former is not finite, the same holds for the latter.

Data	u (threshold)	ξ	σ
Raw	25k	1.4985	90620
Rescaled	145k	1.5868	497436

[5] In most applications of EVT, only the top 5% (or less) of the observations lie above the given threshold. Our 60% is definitely very large, confirming the idea of a whole distribution in the domain of attraction of a fat-tailed distribution, like the Fréchet (Embrechts et al., 2003).

Table 2: Maximum likelihood estimates for the parameters of the GPD approximation of the right tail of the distribution of war casualties. All estimates are significant with a type I error of 5%.

As said, these results are robust to missing or misspecified data. As shown in Cirillo and Taleb (2016), using tools like bootstrap, up to 20% of the observations (in the tail or not) could change, without affecting the results of the analysis.

But if the mean of the distribution of war casualties is not finite, then it means that the tail risk of armed conflicts is not finite. In other words, we could experience any second a single event annihilating humanity. A nuclear holocaust, or even worse.

But can this really be the case?

When dealing with tails, extreme value theory is the right approach. It would be wrong and highly misleading to approach tails using other techniques, mainly relying on normality. However, when using EVT, it is extremely important to take into consideration the real nature of data. Can a conflict kill more than the whole world population?

The answer is clearly no. And this fact needs to be taken into account, if we do not want to be fooled by data.

The distribution of war casualties is necessarily bounded: we surely cannot kill a negative amount of people, but, on the other side, we cannot kill more than the whole world population (at present, 7.3 billion people, according to the United Nations). From a statistical point of view, boundness has one important implication: all the moments of the distribution need to be finite, thus including the mean and the variance. These are the shadow moments, in the terminology of Taleb and Cirillo (2015), i.e. moments that cannot be correctly inferred from data, unless we take into consideration the existence of an upper bound, and we correct for it.

Using the so-called dual distribution (Cirillo and Taleb, 2016), that is a particular log-transformation of the original data, to map them on the bounded support, one can obtain the actual moments of the distribution of war casualties. These moments are naturally finite, but they tend to be much larger than those one could estimate from data using their simple empirical counterparts (which are therefore not reliable). For example, using rescaled data, we discover that the tail mean of war casualties above the 1 million threshold is 6.21 million, against a corresponding sample mean of 3.95 million (1.57 times larger). For a threshold of 50 million victims, the sample mean is 28.22 million, while the true (shadow) mean is 67.17 (2.38 times larger).

When dealing with tail risk, another set of important statistics is represented by quantiles. A quantile is the value above which a certain percentage of observations lie. The top 5% quantile is the value above which we can find 5% of all the observations. Table 3 contains the top quantiles of the distribution of war casualties, using the dual distribution approach on both actual and rescaled data. The results are frightening: there is a 5% probability, using actual data, of obseving a conflict generating at least 2,380,000 casualties; and a 1% probability of conflicts with at least 26.8 million victims. Even worse: our data also support a 0.1% probability of a war killing something like 800 million people, more than 10% of the whole world population. These figures are even scarier when we use rescaled data.

% above	Raw Data $\times 10^7$	Rescaled Data $\times 10^7$
5%	2,380,000	15,400,000
1%	26,800,000	198,200,000
0.1%	801,100,000	4,751,500,000

Table 3: Top quantiles for the distribution of war casualties, as obtained via the dual distribution.

Conclusion: Is there any trend?

The short answer is no.

Our data do not support the presence of any particular trend in the number of armed conflicts over time. Humanity seems to be as belligerent as always. No increase, nor decrease.

Naturally we are speaking about the type of conflicts for which we have performed our analysis, that is to say the largest and most destructive ones. We cannot say anything about small fights with a few casualties, since they do not belong to our data set –however it is crucial that, as a central property of the fat-tailedness of the process, a decline in homicide does not affect the total properties of violence and anyone's risk of death. As we said, the mean is tail driven.

At the best of our knowledge no available data set contains enough information to make credible statements about statistically significant trends in the number of conflicts over time, unless we really think it is reasonable to extrapolate long-term trends on the basis of sixty years of observations, like those after WW2. Given the inter-arrival times we have observed above, it would be quite naïve to act that way.

If we focus our attention on our data set, and in particular on the observations belonging to the last 600 years (from 1500 AD on), for which missing observations should be fewer and reporting errors smaller, our analyses suggest that the number of large conflicts over time follows a homogeneous Poisson process. In a similar process, the number of observations over time, once we fix a given time interval (say 50 years), follows a Poisson distribution. The number of expected data points only depends on the length of the time interval we choose. For intervals of the same size, the expected number of observations is the same, because the intensity of the process does not vary over time. In simple terms, this finding supports the idea that wars are randomly distributed accidents over time, not following any particular trend, as already pointed out by Richardson (1960).

Interestingly, similar conclusions can also be derived by a simple descriptive analysis of data, something that should make our results accessible also to non-statisticians.

In Figure 7 and 8 we show the average number of casualties and the number of conflits in the period 1500-2015, using a non-overlapping moving window of 20 years. This means that

both the average casualties and the number of conflicts are computed for the periods 1501-1520 (there is no observation in 1500),1521-1540,1541-1560 and so on.

In Figure 7, the red dots represent rescaled data, while the green ones are the actual observations. It is quite evident that, for what concerns the average number of casualties, no clear trend is observable.

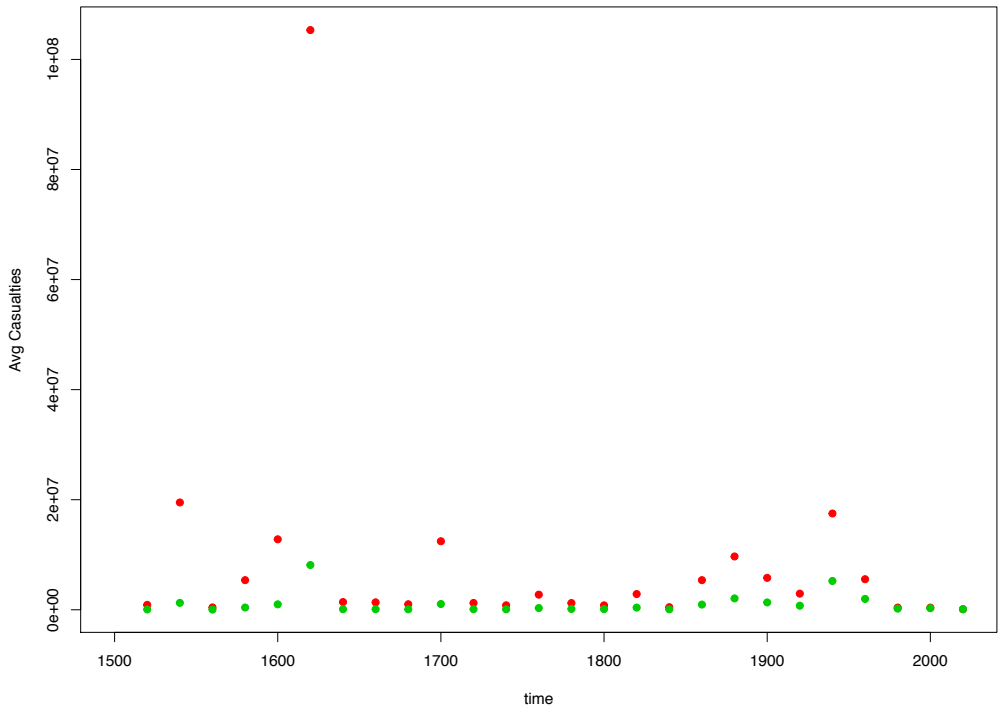

Figure 7 Average number of casualties in the period 1500-2015 using a non-overlapping moving window of 20 year. Red dots: rescaled data. Green dots: actual data.

In Figure 8, we show the number of armed conflicts in the same moving windows as per Figure 7. In this case, the number of conflicts seems to be increasing over time, even if the volatility itself appears to be higher. This is an interesting phenomenon from a statistical point of view, as it makes the simple inference based on a few years of data (namely the last 60 ones, as in the "Long Peace" theory) not at all reliable. While we are aware that this behavior could be due to a historiographical reporting bias, according to which more recent conflicts are more likely to be recorded in data, we would like to stress how, in any case, no Long Peace is observed. In particular, the last 200 years prove to be quite belligerent and "stable".

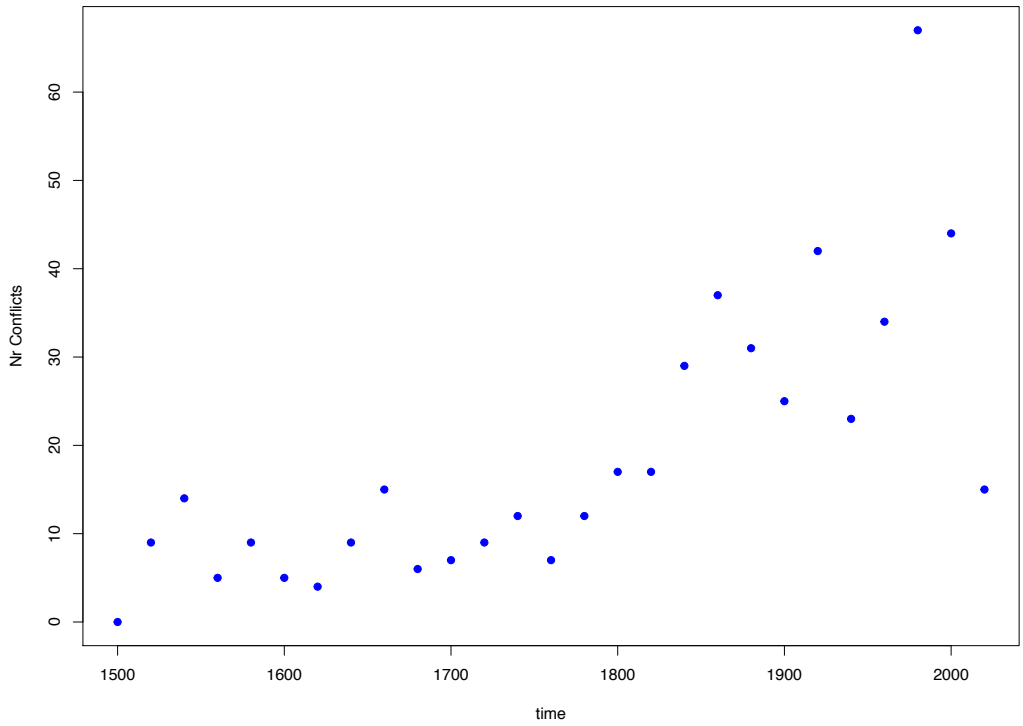

Figure 8 Number of conflicts between 1500 and 2015, using a non-overlapping 20-year moving window. The number of conflicts is clearly increasing over time, together with their volatility.

One final comment on the finer-grained period since WW2. Figure 9 illustrates the mistake made in theorizing about what has happened since 1945. Aside from the fact that we are picking a spike and that a drop is natural after every spike. The way to properly look at the issue is to consider the entire history of wars, simulating from the process, and checking how many periods do not look like the stretch since 1945 –(by generating simulated regression coefficients). Alas, we are about .37 standard deviations away from absence of trend. Note that we are ignoring the survivor bias.

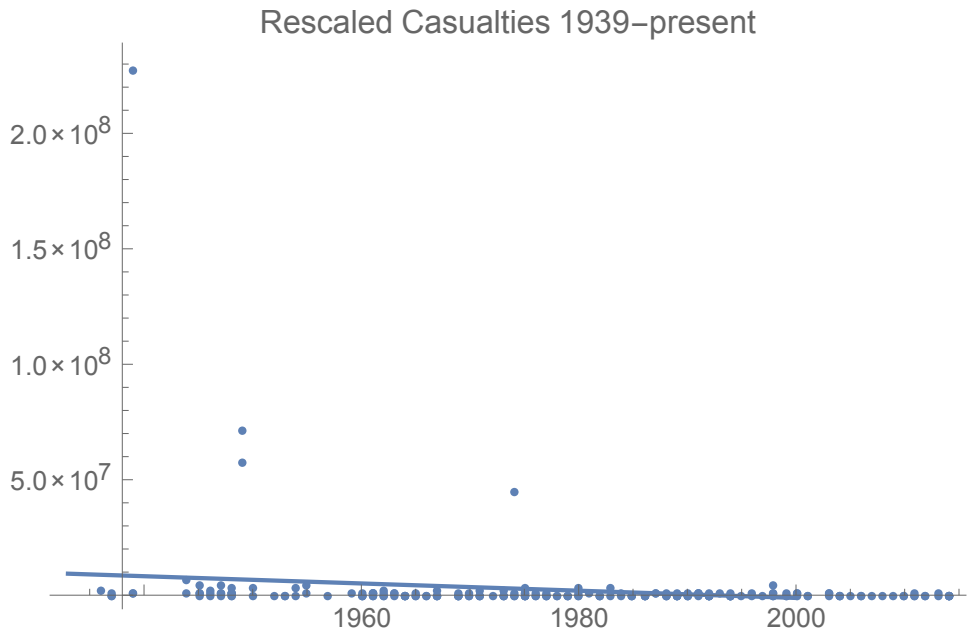

Figure 9 Violence drop since 1945. Divergence from the process to call it a "trend" is patently not statistically significant, .37% of a standard deviation away. No scientist builds a theory from .37 standard deviations.

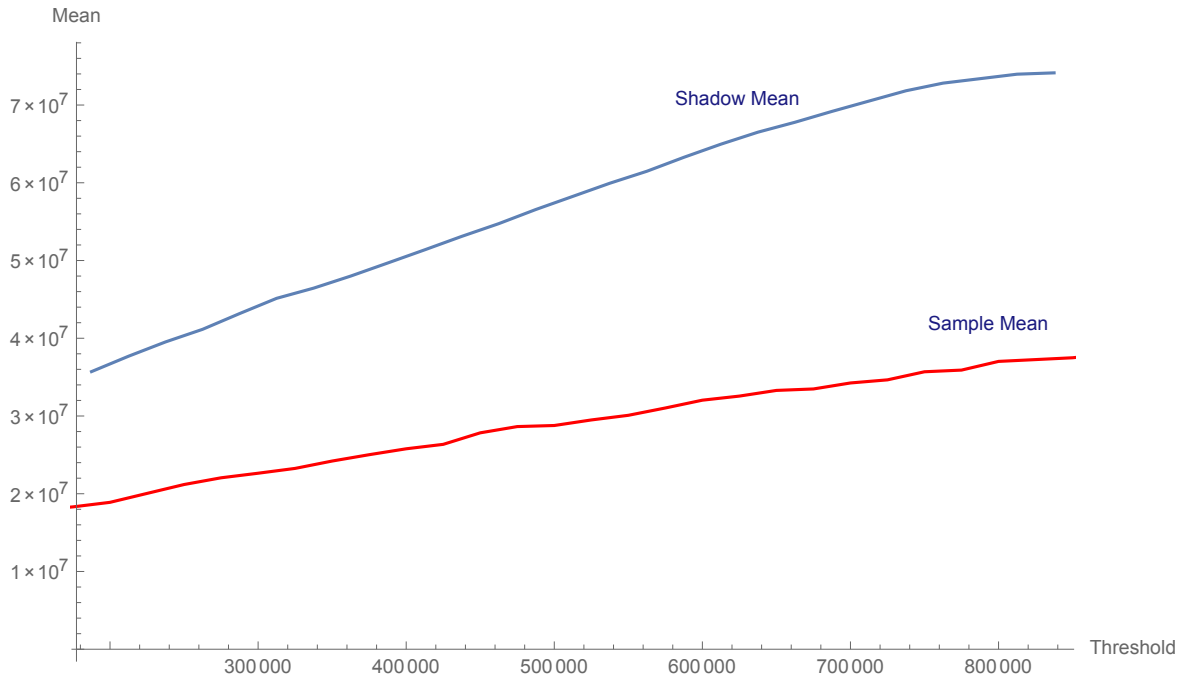

Figure 10 Shadow mean at different threshold.

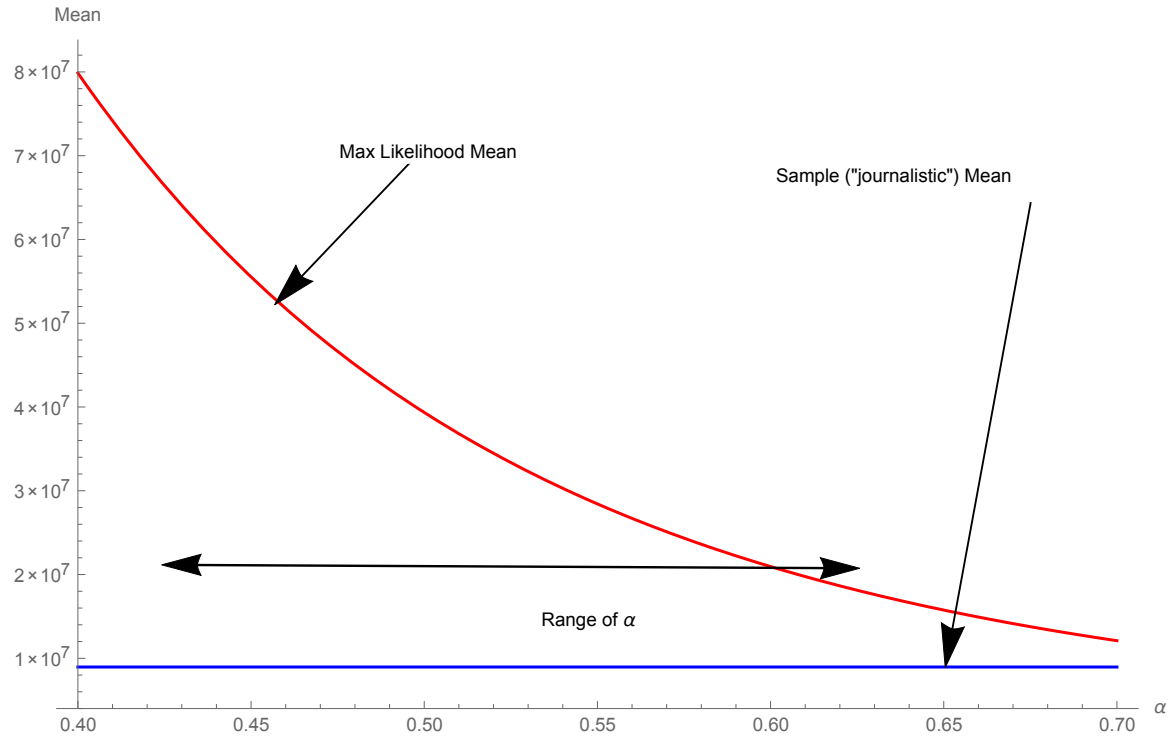

Figure 11 Sample mean (journalistic) and maximum likelihood mean at different values of the tail exponent. We notice that a drop in the tail exponent causes an asymmetric effect , hence uncertainty and lack of precision about the tail means worse mean (Taleb 2016).

The fact that the "shadow mean" or maximum likelihood mean is in excess of the sample mean across potential values of the parameters (Figures 10 and 11) is not trivial. It means that even a "real" statistically significant drop in violence would not alter by much the gravity of the situation: the world is even more dangerous than it looks.

We connect with the rest of the chapters as follows. Our paper may add some arguments or take sides into the democratic peace debate entailing Russett (2017, this volume) and Gowa and Pratt (2017, this volume), as follows: there has been, at the macro level, a rise in democracy, but no evidence of decline of wars and general violence. It is always tempting to assume that the rise in institutions has contributed to change in the structure of the world – just as many arguments have been made that the creation of the federal reserve and other financial institutions have contributed to stability. In finance, this argument turned out to be wrong: extreme events have been at least as severe (and if anything have risen) in spite of the development of such institutions. It may be the same with violence.

References

Balkema, A.A., de Haan, L. Residual life time at great age. The Annals of Probability 5, 792-804 (1974)

BBC (2012). In our times: The An Lushan Rebellion. http://www.bbc.co.uk/programmes/b01by8ms.

Berlinski, D. *The Devil's Delusion: Atheism and its Scientific Pretensions*. Basic Books, New York, (2009)

Buckle, H.T. *History of Civilization in England*, Vol. 1, London: John W. Parker and Son (1858)

Cederman, L.E. Modeling the size of wars: from billiard balls to sandpiles. The American Political Science Review 97, 135-150 (2003)

Cirillo, P. Are your data really Pareto distributed? Physica A: Statistical Mechanics and its Applications 392, 5947-5962 (2013)

Cirillo, P., Taleb, N.N. On the statistical properties and tail risk of violent conflicts. Physica A: Statistical Mechanics and its Applications 452, 29-45 (2016).

Clauset, A., Woodard, R. Estimating the historical and future probabilities of large terrorist events. The Annals of Applied Statistics 7, 1838-1865 (2013)

Contestation. Journal of Conflict Resolution 53, 934-50 (2009)

Cox, L.A., Greenberg, M.R. (eds.) Special Issue: Advances in Terrorism Risk Analysis. Risk Analysis, online ISSN 1539-6924, (2013)

de Haan, L., Ferreira, A. Extreme Value Theory: An Introduction. Berlin, Springer Series in Operations Research and Financial Engineering, 2006

Durand, JD, "The population statistics of China, AD 2 – 1953," *Population Studies* (1960), Vol. 13, No. 3, p.209,223

Embrechts, P., Klu¨ppelberg, C., Mikosch T. Modelling Extremal Events. Springer, Berlin,

Epstein, R. Book Review: The Better Angels of Our Nature: Why Violence Has Declined. Scientific American, October 7, (2011)

Falk, M., Hüsler, J., Reiss, R.D. Laws of small numbers: extremes and rare events, Birkhäuser, Berlin, (2004)

Fitzgerald, Charles Patrick, *China: a Short Cultural History,* Cresset Press,,1935

Friedman, J.A. Using power laws to estimate conflict size. Journal of Conflict Resolution 59, 1216-1241 (2015) 35

Gnedenko, D.V. Sur la distribution limit´e du terme d'une s´erie al´eatoire. Annals of Mathematics 44, 423-453 (1943)

Goldstein, J.L. Winning the war on war: the decline of armed conflict worldwide. Penguin, New York, (2011)

Gumbel, E.J. *Statistics of Extremes*. Cambridge University Press, Cam- bridge, (1958)

Hayes, B. Statistics of Deadly Quarrels. American Scientist 90, 10-14 (2002)

Horne, A. *A Savage War of Peace.* New York Review Books Classics, New York, (2006),34

Inkester, N. (ed.) *Armed Conflict Survey* 2015. IISS, London, (2015)

Kleiber, C., Kotz, S. *Statistical Size Distribution in Economics and Actuarial Sciences*, Wiley, New York, (2003)

Klein Goldewijk, K., van Drecht, G. HYDE 3.1: Current and historical population and land cover. In Bouwman, A.F, Kram, T., Klein Goldewijk, K. (eds.) Integrated modelling of global environmental change. An overview of IMAGE 2.4. Netherlands Environmental Assessment Agency, The Hague (2006)

Mandelbrot, B. B. -*The Fractal Geometry of Nature*, Freeman5 (1983)

Maronna, R., Martin, R.D., Yohai, V. *Robust Statistics - Theory and Methods*. Wiley, New York, (2006)

Mueller, J.E. Retreat from Doomsday: the obsolescence of major war. Basic Books, New York, (1989)

Necrometrics, (2015) Available at: http://necrometrics.com.

Norton-Taylor, R. Global armed conflicts becoming more deadly, ma jor study finds. The Guardian, Wednesday 20 May, (2015) Available at: http://www.theguardian.com/world/2015/may/20/ armed-conflict-deaths-increase-syria-iraq-afghanistan-yemen

Phillips, C., Axelrod, A. *Encyclopedia of Wars* 3-Volume Set. Infobase Pub- lishing, New York (2004)

Pickands, J. Statistical inference using extreme order statistics. the Annals of Statistics , 1, 119-131 (1975)

Pinker, S. *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Press, New York, (2011)

Pinker, S. , Taming the Beast Within Us, Nature, 478, 309-311(2011b)

Reiss, R., Thomas, M. Statistical Analysis of Extreme Values, Birkh¨auser, Berlin, (2001)

Richardson, L. F. Variation of the Frequency of Fatal Quarrels with Magnitude. Journal of the American Statistical Association 43, 523-46 (1948)

Richardson, L.F. Statistics of Deadly Quarrels. Quadrangle Book, Chicago (1960)

Scharpf, A., Schneider, G., Noh, A., Clauset, A. Forecasting of the risk of extreme massacres in Syria. European Review of International Studies 1, 50-68 (2014)

Seybolt, T.B., Aronson, J.D., Fischhoff, B. (eds.) Counting Civilian Casualties, An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict. Oxford University Press, Oxford, (2013)

Taleb, N. N. -*The Black Swan: The Impact of the Highly Improbable*. Random House and Penguin, (2007/2010).

Taleb, N.N. *Silent Risk* (2016). Available online: www.fooledbyrandomness.com/FatTails.html

Taleb, N.N., and P. Cirillo. "On the shadow moments of apparently infinite-mean phenomena." *arXiv preprint arXiv:1510.06731* (2015).

United Nations - Department of Economic and Social Affairs. 2015 Revision of World Population Prospects. UN Press, New York, (2015)

United Nations - Department of Economic and Social Affairs. *The World at Six Billion*. UN Press, New York, (1999)

van der Wijk, J. Inkomens- en Vermogensverdeling. Publication of the Ned- erlandsch

Viertl, R.. *Statistical Methods for Non-Precise Data.* CRC Press, Boca Ra- ton, (1995)

Villasenor-Alva, J.A., Gonzalez-Estrada, E. A bootstrap goodness of fit test for the generalized Pareto distribution. Computational Statistics and Data Analysis 53, 3835- 3841 (2009)

Wallensteen, P., Sollenberg, M. Armed Conflict 1989-2000. Journal of Peace Research 38, 629-644 (2001)

White, M. *The Great Big Book of Horrible Things*. W.W. Norton and Company, New York (2011)

White, M. *Atrocities*. W.W. Norton and Company, New York, (2013)

Wolfram, S. *A New Kind of Science*, Wolfram Research (2002)