Unfolding History: Classification and analysis of written history as a complex system

D. Massad[1], E. Omodei[2], C. Strohecker[3], Y. Xu[4], J. Garland[5], M. Zhang[6], and L. F. Seoane[7,8]

[1]Department of Computational Social Science George Mason University, Fairfax, Virginia, USA.
[2]ISC-PIF, LaTTiCe-CNRS, cole Normale Suprieure, Paris.
[3]Vice Provost, Academic Affairs, Rhode Island School of Design.
[4]Department of Physics, Florida State University, Tallahassee, United States.
[5]Department of Computer Science, University of Colorado, Boulder, Colorado, USA
[6]Center for Complex Systems and Brain Sciences, Florida Atlantic University
[7]ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Barcelona, Spain.
[8]Institut de Biologia Evolutiva-CSIC, Barcelona, Spain.

Sep. 16, 2013

1 Introduction: written history as a complex system

Before writing, chronicles were told by word of mouth. They conveyed information of important historic events that could readily be mixed up with legends – or become them. Chronicles needed to be easily accessible for people to pass them along, and stories lived on in listeners' short memories. This collective, ever-waning awareness was the technology that kept the tales extant, and this technology required a continuous effort in reminding people never to forget. This way of keeping stories alive also imposed important constrains on the kind and form of the narrative material. Breathtaking epics were perhaps more likely to survive, which could easily tend toward mystification of pre-historic characters. These epics had to be kept in the form of repetitive, easy-to-learn patterns. It comes as a pleasant surprise for us that diverse ancient poems take similar

forms in different cultures, that they present similar – if not identical – rhythms, verses, and topics as if to make their transmission easier, and that these similarities can arguably be attributed to the biological reality of human beings [4]. In other words, the available biological technology for the preservation of tales has imposed important constrains on what and how pre-historic events can be produced and preserved.

As writing began to emerge, human actions began to be registered in a physical substrate offering new possibilities and imposing new constraints. The technology of *writing* is not definitive, but evolves through a tinkering of techniques from handwritten words, to the printing press to newer devices that are constantly evolving and being invented. Yet human history remains registered in written form. Similar to the restrictions we can see on *sung* stories, the written word has its own rules, which we suggest becomes reflected in what historic events are record and moreover how they are recorded.

In this paper we undertake the task of framing written history as a complex object worthy of the attention of the complex systems research community. We bring to light and emphasize written history's complex interactions with social dynamics, language, and–of course–the historic events that constitute its very essence; whose parts and distinct interactions are almost impossible to distinguish. We argue that written history deserves attention by the complex system community, following the recent trends of culturomics, which conceive human culture and its evolution and dynamics as a complex system readily accessible for our study [13].

Here we make a distinction between *objective historic events* and *written history*. We assume that the former have once taken place and that the later accounts for the former, but perhaps not in it's entirety and perhaps things have been added that didn't actually happen. In this way, written history becomes a record of human activity that may or may not have occurred [27]. Written history departs from objective historic events, in what can be viewed as a complex process in it of itself and once this diversion occurs reconstruction can only be accomplished through the subjective account of peasants and scholars [10]. In this departure from objectivity, true reality vanishes and written history becomes the new reality and the only reference to understanding the past. As a result we must question how faithful to the original events written history actually is–if faithful at all. This makes us wonder how a true a definitive version of history could be assembled, if such a definitive version exists, and how individual contributions would increase the knowledge–in terms of rigorously defined information–of this event. A particularly interesting case to examine would be when information or paradigm shifts occur which cause the way we as a society view a historic event, and how this new information effects the rewriting of that particular event.

In an ideal world, we would turn our attention to history books and journals and while this might not be a chimera anymore thanks to diverse technological developments [13], it is still a daunting task. And such data acquisition was out of scope given the time constraints of the summer school. As such we needed to turn our attention to a more readily available account of history and as such we

chose Wikipedia as a testbed for our tools. With Wikipedia we have access to the record of edits of each historical event (wiki page), together with plenty of meta information; it is a massive trove of data, written by society itself, whose paradigm shifts may affect the way history is written. However it has serious drawbacks as well, notably, Wikipedia has only existed for a short period of time in comparison to written history and as such some of the phenomenons that we wish to research may not be present yet. Furthermore, Wikipedia has its own internal dynamics regarding article growth, fame of editors and editor contributions [1, 2, 3, 11, 22]; this in turn might shadow the very subtle processes that we wish to study.

As motivation for our proposal, we now discuss a few phenomena or dynamics that actually have been observed in written history. To aid this discussion we use metaphors from biology and ecology, but intend these as similes only. In Sections 2 and 3 we explain a few rudimentary tools borrowed from statistics, information theory, and natural language processing; and we apply these tools to toy examples in order to illustrate what features we could render if the methods were put to work in a rigorous and systematic manner. A reflection about our work and future lines of research follows in section 4.

1.1 Written history as a data-mining of historic events

Our simile for this section is the genome. DNA stores information about the developmental process of the species. It arises very early and seems common to all forms of life on earth. It constitutes a major transition in evolution [14] that provides living beings with a technology capable of information storage and of information mining through the evolutionary process. Indeed, a recent argument asserts that the information captured by a population from the environment is maximized through natural selection [12].

In a similar way, we propose to look at written history as a physical recording of events that have happened, but a recording in which everything cannot fit. Because of technological limitations and other impediments perhaps stemming from social or psychological appreciations, perhaps because of ideological reasons [13], what enters written history is constrained. The information that is considered necessary in order to keep a record of human activity is somehow limited.

In the case of DNA and natural selection, the proper metrics seems to be Fisher information [12]. However, we do not know what would be the right approach to written history. In fact, we do not even know whether any process leads to an optimal information transfer from reality into written history, as natural selection arguably does for DNA; and we do not know what factors ultimately restrict the size of written history–if any. However, we intend to apply a rigorous information theoretical approach to this problem in the future:

- We can compute proxies for the Kolmogorov complexity of texts [16, 6]. A procedure for that—and one that also connects with the topics during the CSSS [7, 8] – would be to infer an ϵ-machine that would capture the

intricacies of the string that a text is made of. More approaches exist to this problem, some of them severely criticized [20, 26].

- Using Wikipedia articles for example, we can study transfer information [22] between successive versions of the same article and try to understand how new information is continuously incorporated. If we had access to a great corpora of historic books we could analyze how different points of view add up information about a general fact.

- Turning our attention once more to Wikipedia, if we analyze current/ongoing events (as illustrated in section 3) we hypothesize that we could begin to understand how information relevant to a society is generated in real time.

We must remark the highly speculative nature of this methodology. It was our intention to elaborate a little bit on this idea and put it to work on some examples. This was however not possible due to time constraints, but the idea is still appealing to us and we considered it worth communicating in this report.

1.2 An ecosystem of topics

Another metaphor assisting our inquiry is ecology and its techniques, which are applicable at multiple levels. We begin with the lowest one possible. When written for a broad public, a text generally must obey some consistency rules, even if they are implicit. The text usually is required to be coherent in order to be effectively informative; an incoherent text may be rejected by the readers. We hypothesize that these constraints should leave some imprint in any text – not only historic texts – and that these imprints could be measured. The constraints could introduce serious limitations regarding what topics fit together and how the length and purpose of a text may limit the number of topics it can address.

Furthermore, relationships between events and people, between records and references to prior recordings, and between words in a single record all may contribute to a narrative's completeness, coherence and comprehensibility. Just as the medium and the recorder are limited by physical, social and psychological constraints, so is the reader who reconstructs meaning from the text with an individual perspective.

At a higher level, while narrowing the discourse strictly to historic texts, we find that events of human history are very dependent on each other. Different circumstances converge to trigger a burst of historic events in a causal manner, and prominent characters may play important roles through the course of their lifetimes. Salient dates might also help to trigger new events. All these phenomena are worth researching. Interrelationships between different historic events and figures is made visible thanks to Wikipedia, which provides hyperlinks between pages devoted to important historic moments and actors [3].

We hypothesize that some of the consistency-related rules outlined above might have effects at the Wikipedia network level, with edits in one page spreading over the network, perhaps to create greater coherence within the global picture. To study such effects, we apply the methods explained in more detail in

Section 2. These methods utilize the time series of edits of different Wikipedia pages: the correlation is calculated between pairs of these time series, indicating whether there is a tendency to edit different pages at similar times or not. The results from these calculation are then compared to a null model which tells us whether the correlations found are significant. This method finds two weaknesses:

1. the null model – derived from a bootstrapping of the data –may not be adequate

2. while we can argue that our data is significant regarding the null model, it is extremely difficult to argue that the significant correlation stems from the kind of procedures that we want to analyze

We must reserve clarification of both of these troubling issues for future work, as they are both outside the scope of the current project.

A most daring vision of history as an ecosystem would see its different affairs as individuals seeking food – i.e., attention by human authors who would seek to improve the quality of different parts of written history [9]. This analogy would connect tangentially with the ideas in section 1.1: it is likely that the most important topics receive more attention by authors, highly conditioning where new information is allocated.

2 Investigating correlation networks as a tool for unfolding history

In this section we describe a methodology that we found interesting and one we believe is promising for investigating this problem in the future. Qualitatively, we obtain the edit records of a collection of Wikipedia sites. These records can have up to a *minute* precision, but as we will show, using such a small time scale introduces problems and so we use longer time scales for our calculations. From the edit histories (time series) we build a weighted network based on the correlation of the edit records of different wiki sites. This is a picture of *how likely* it is to edit couples of concepts at a similar time. Correlation was used to provide a first idea of the general approach, notwithstanding the known problems of this coefficient [19]. We note that instead of this measure we could be using mutual information or information transfer among others.

Because Wikipedia is so large, we only work with a small sample at any given time. To collect these records we chose a *seed page*: a starting event or topic. From this seed page we derive a *depth-one* network consisting of all the pages the seed page links to. We extract the edit history for these pages, choose a time scale (day, week, month) and compute the number of edits per period. This provides us with a set of N time series $X \equiv \{x_i(t)|t = 1, ..., T, i = 1, ..., N\}$ where N is the number of pages in the network, including the seed and T is the number of time clicks since the seed page was created; in words $x_i(t)$ is the

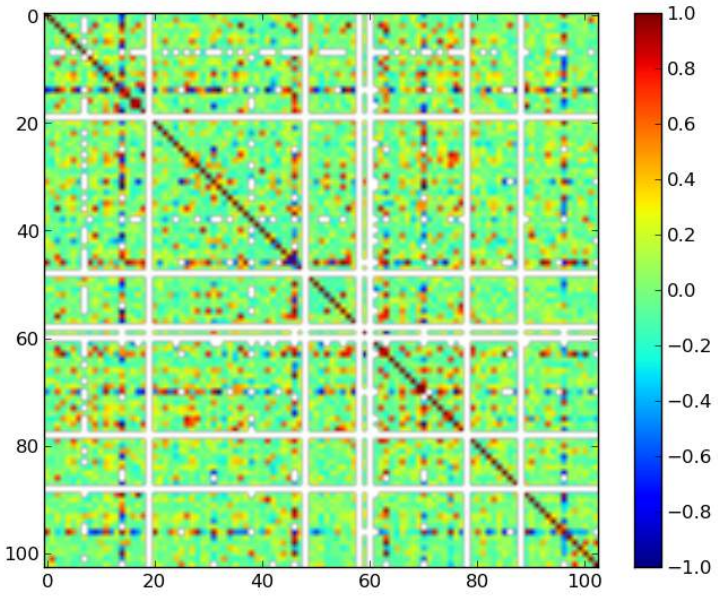

Figure 1: **Correlation of the edit histories in the network derived from the Wikipedia article *2012-13 Cypriot financial crisis*.**

number of edits of the i-th page during the t-th time period of the chosen unit– usually days. Of course, a page does not need to exist for the whole Wikipedia life span. This implies that different time series will have different lengths, which may or may not overlap. We believe it only makes sense to calculate correlations between *overlapping* time stretches. In figure 1 we show the correlation matrix for the depth-one network derived, using the seed page *2012-13 Cypriot financial crisis* from Wikipedia.

In Figure 1 we observe that the edit history of different wiki pages present different levels of correlation, spanning between -1 and 1. It is tempting to build a correlation-based network where each page is a node and those pages correlated above a threshold r_θ are linked. As an example, in Figure 2 it is shown the $r_\theta = 0.5$ network derived from the *Berlin Wall* Wikipedia page in English.

As noted, it is very tempting to take this approach; but we must assess the question of *how significant are the correlations that result.* Indeed, for a collection of time series, each of which consists of an independent random variable sampled several times, we expect correlations arising by chance, although the variables should not be correlated *in average.* We need to eliminate those correlation that arise by *pure chance* from our data. To this end we used a very

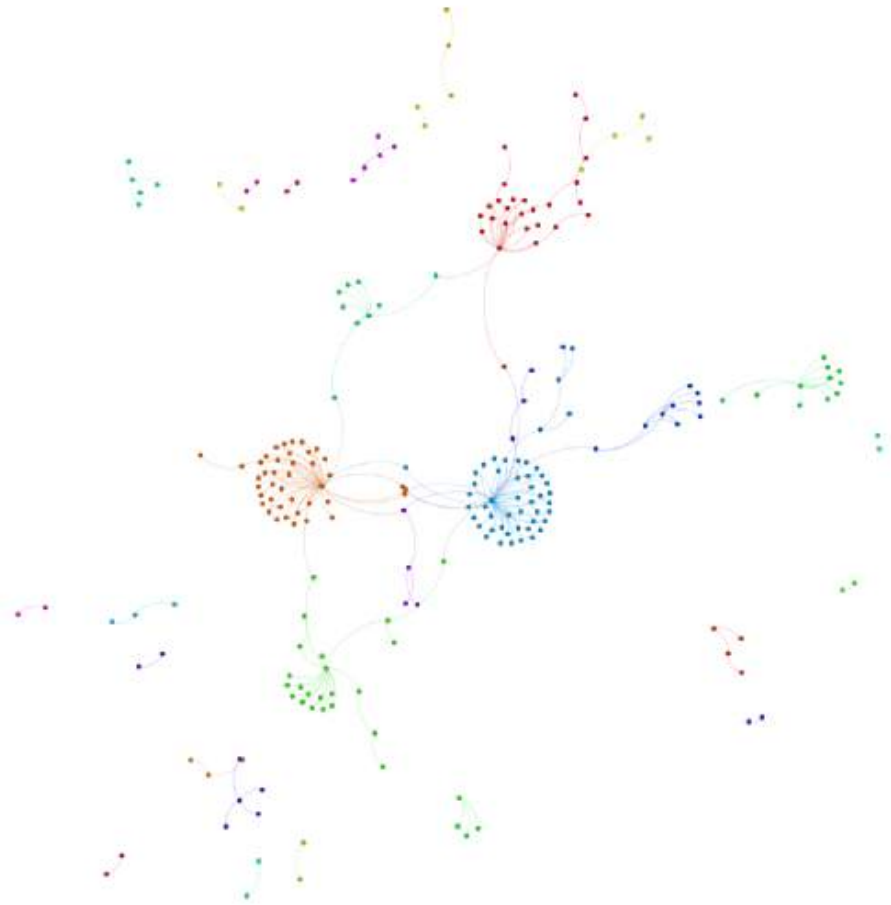

Figure 2: **Network for Berlin Wall with $r_\theta = 0.5$.** A network was derived following the hyperlinks from the English *Berlin Wall* page of Wikipedia, computing the correlation between edit histories of different pages, and choosing only those pages whose correlation is above r_θ. We can observe clustering, which we believe is due to pages focusing around a similar topic, but this needs to be be rigorously tested.

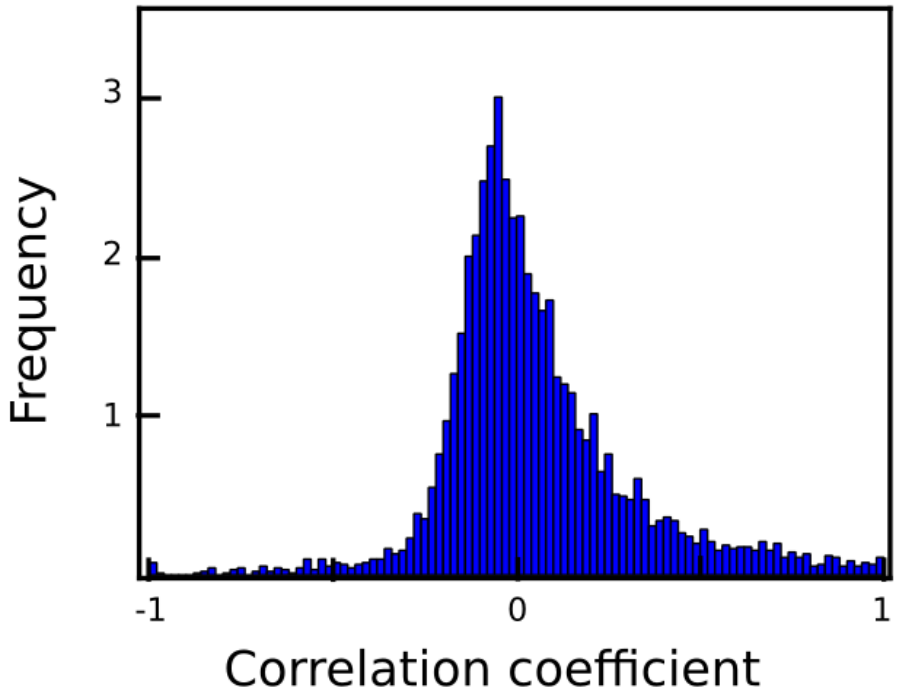

Figure 3: **Correlation histogram derived for the *2013 Cyprus crisis.*** From this histogram is excluded autocorrelation, which amounts to 1 and biases the visual presentation. We appreciate that the distribution is asymmetric with a heavier tail at the right, suggesting that the pages are *on average* positively correlated. We must assess whether or not this slight asymmetry is significant.

rudimentary bootstrapping.

To being, we made a histogram of the correlations that show up in a given network, as illustrated in figure 5. From that histogram we recorded the bins. We proceeded to shuffle every time series, meaning that for each $x_i(t)$ we assign $\tilde{x}_i(t) = x_i(t'_i)$ where t'_i samples the domain of t in a random manner and without repetition. This way we generate a synthetic data set $\tilde{X}$ from where we extract correlations, as we did from the original time series. We can repeat this procedure to get as many synthetic data sets as desired; say N_B, then $\{\tilde{X}_j | j = 1, ..., N_B\}$. These shuffled data constitute a null model against which to compare our data. This null model could be refined, as discussed below.

For each $\tilde{X}_j$ we compute the frequency with which correlations arise within each bin of the original histogram. From all the synthetic time series we can estimate the average frequency with which data from the null model present correlations within a given correlation bin. We can also estimate the standard deviation of the frequency expected at each bin. In Figure 6 we plot the same

histogram as in Figure 5 (derived from the 2013 Cyprus crisis wiki page) where it has been superimposed with the average frequency expected at each bin after bootstrapping, and the average plus or minus three times the standard deviation. Assuming the frequency within each bin follows a Gaussian distribution, the likelihood that a bin presents a frequency outside the boundaries of three standard deviations by chance alone is less than ~ 0.0027, and the probability to find that number strictly above the boundaries is less than ~ 0.0014.

For the network derived from the Cyprus crisis, we observe that in the original data very few bins significantly deviate from what would be expected purely by chance from the null model (Figure 6), even when the null model is a very rudimentary one which was designed to break apart any correlation within each time series. Remarkably, larger correlations are not more significant than lower correlations, meaning that when plotting a network whose nodes are correlated above a threshold r_θ—as we did in Figure 2—it is very likely that we are plotting randomly correlated time series.

It is very interesting that the null model retains some of the features of the original distribution such as not being centered around 0 and not being symmetric, even when it derives from such a crude bootstrapping. Note that these features could stem from deep characteristics of the time series—e.g., most of the time there are not edits to an arbitrary Wikipedia page.

We could elaborate more complicated null models to compare with. For example, when shuffling the data we did not care that the activity in a wiki page during one day might be strongly correlated with the activity the following day—we just tore apart any correlation occurring within a particular time series . Indeed, if the page reports ongoing events—as in the Cyprus crisis—we expect that events happen after each other and new information should be available to append to Wikipedia one day after the other. One refinement would be, for example, to generate a null model that shifts the origin of the whole time series. This should be implemented in future work.

Now the good news: In Figure 7 we plot the result of this procedure for several historic events such as World War I and II and the recent Eurozone crisis, as well as for historical concepts and figures such as the Roman Empire, the Berlin Wall, and George W. Bush. In all these cases we observe how the proportion of pages that present a correlation within a range of $\sim (0.15, 0.65)$ is significantly larger than expected from the null model. This indicates that there are some processes—that we discuss below—that moderately increase the simultaneous editing of groups of related historic events and figures; but these processes *do not* strongly increase the observed correlations. Furthermore, the amount of *moderately more correlated pages* is significant in comparison with our null model.Remarkably again, larger correlations seem to be within the range expected by the null model.

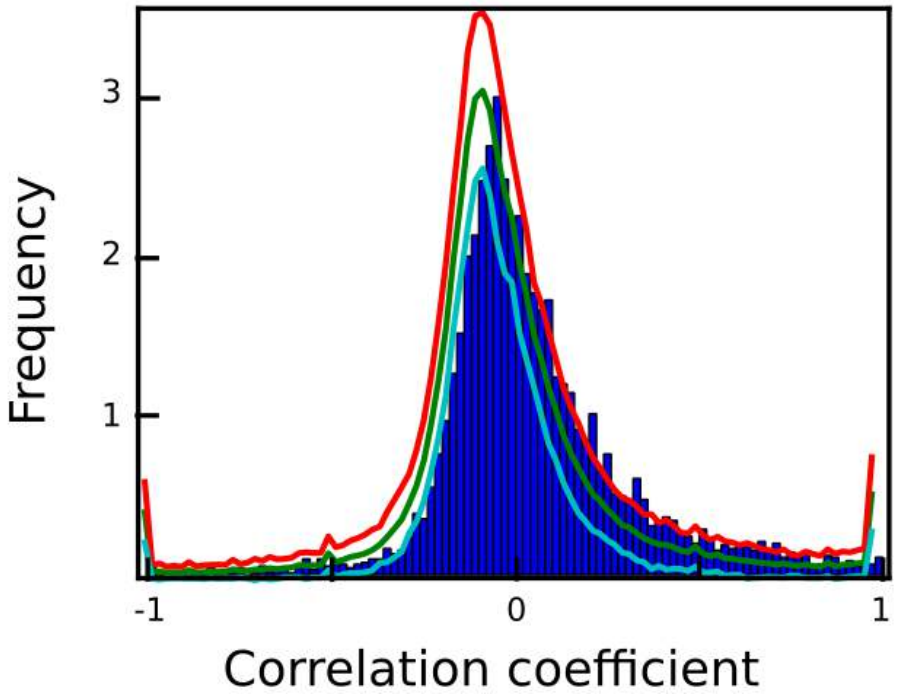

Figure 4: **The bootstrapping provides a null model.** Here we plot the histogram of the correlations from Figure 5 (derived from the *2013-12 Cyprus crisis*), as well as the average frequency expected in each bin (green) given the null model produced by bootstrapping, the average minus 3 standard deviations (light blue), and the average plus three standard deviations (red). The probability that we observe a frequency outside the three standard deviations boundaries by chance alone is $\lesssim 0.0027$. For this example, there are few bins whose deviation from the null model is significant, meaning that the observed results (asymmetry in the correlations, network of highly correlated items) could have happened just by chance.

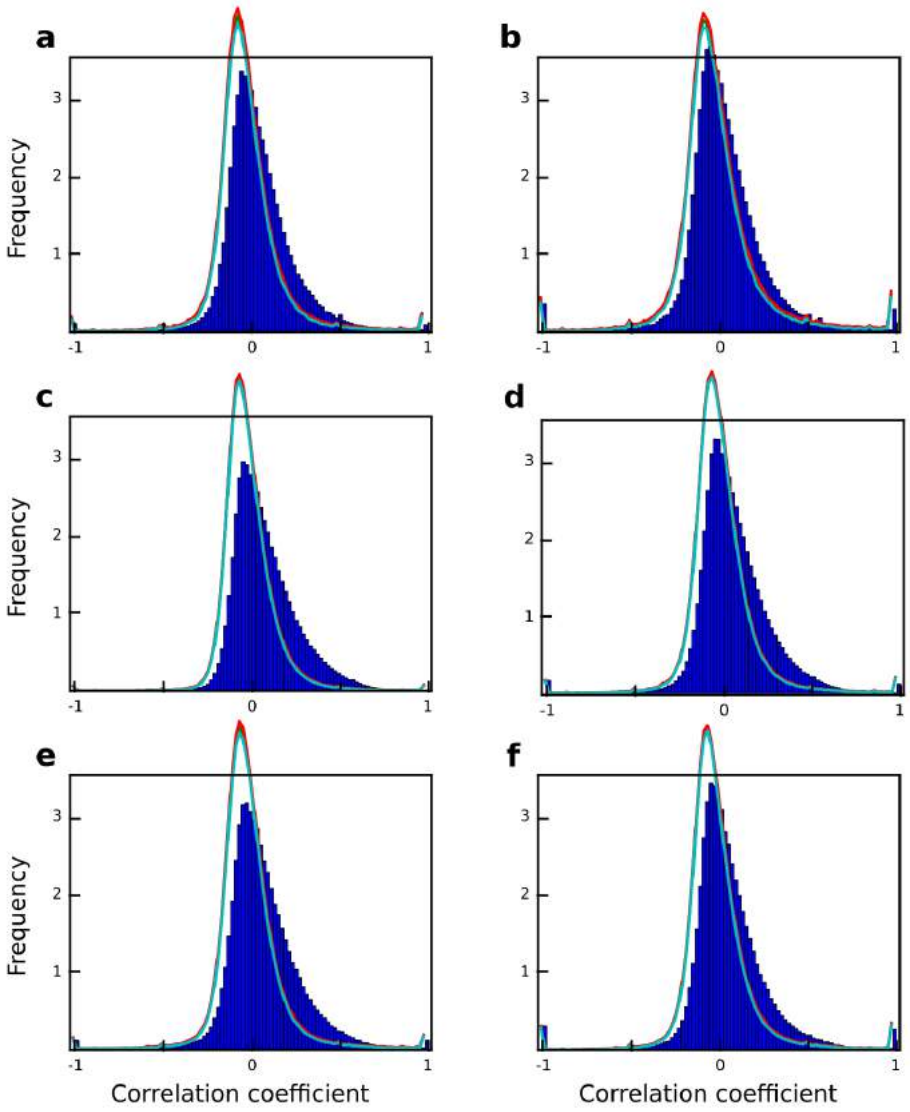

Figure 5: **Data and bootstrapping for several historic figures and events.** **a** Berlin wall, **b** Eurozone crisis, **c** George W. Bush, **d** Roman Empire, **e** World War I, **f** World War II. When compared to the bootstrapped synthetic distributions (superimposed curves of different colors) we find that the networks derived for all these Wikipedia articles feature a significant increase in pairs of sites that are more highly correlated than we expect purely by chance. Let us note that the increase in correlation is not significant for one particular pair of sites, what is significant is how many pages are correlated more than the null-model predicted. This might be indicative of subtle waves of edits spreading throughout the Wikipedia network.

2.1 Distinguishing correlation significance on causation in written history

While we can identify correlations between time series of Wikipedia edits, we are very far from showing the causes for those correlations. Let us revisit some possibilities.

The most appealing—and difficult to prove—hypothesis is precisely the one we seek: that correlations arise from the inner dynamics of written history. Indeed, it is fair to assume that all pages linked to from another one have got some kind of conceptual relationship (since they are linked by a common site). Because we observe a significance of the simultaneous editing of such a corpora of related pages, we wish to argue that they are being edited because a coherent version of written history is being forced upon the historic accounts. This would be one mechanism by which the observed correlations could arise.

We must also acknowledge that Wikipedia has its own dynamics regarding the growth of an article and others related to it. These do not necessarily attend to prospective internal rules of written history. Such dynamics might be related to interactions between users or rules stemming from the online editing of wiki pages. These might also be worth studying—and have been considered by many [22, 11, 2, 1, 3]—for people interested in, say, social dynamics, collaborative behavior, and such. But these phenomena depart from our interest in finding out intrinsic rules of written history.

Finally, it is very likely that different events, concepts, and people we examined on Wikipedia are explicitly related in the real world. As a trivial instance: the 2013 Cyprus crisis was the reason why the Laiki Bank (also known as the Cyprus Popular Bank) became an important player in the historic scene. It is obvious, then, that talking about one concept is likely to happen at the same time as we write about the other. It is extremely hard to disentangle such real-world, daily correlations from correlations emerging from the written history trying to be consistent with itself. Even more so, such real-time relationships and the precise mathematical way in which information from one page is related to information from the other page is potentially also very interesting for our studies.

3 A semantic approach based on natural language processing

In this section we illustrate another approach to our problem, this time based on semantic analyses of texts that evolve over time. In a nutshell, natural language processing tools – under intense development at the moment – allow us to extract important topics or items from a text. We can consider this as a coarse graining of the semantic information that will enable an abstraction from the very complex system that the text is, and observe how definite and quantifiable features evolve over time.

Wikipedia articles are evolving texts and it is not so clear that they reach a

stable final version – which is an observation of notable interest to ascertain some dynamics of written history. Analyzing what key words are obtained from a text in its successive versions we can get a grasp of what was important over time for the authors – in our case, the Wikipedia community – to describe a given historic event. If Wikipedia had existed over many decades, we could associate the rise and fall of subjects that describe an event with shifts in the society: changes in the paradigm that stress different aspects for the understanding of human condition. We exemplify such a phenomenon below, but in here the time span is relatively short and it seems daring to attribute the observed changes to such a change of paradigm in the society. Rather, and also very interestingly, the raise of a very precise topic might be associated to a tendency of the written history towards a complete and more technical account, as we will see.

The method described below is also very interesting if it were applied to an ongoing historical process. This is illustrated with a brief example towards the end of the section. In the case of ongoing history we expect more variability on the *main topics* since the players are not clear a priori. As stated above, the described method renders an interesting coarse graining; not of a complex text anymore, but probably of a complex situation which is still unfolding. We speculate that this can be an interesting approach to the actual evolution of historic events.

Instead of working with a whole network of interlinked Wikipedia sites, we chose a transcendent event that we wish to analyze. For the selected Wikipedia page we download the whole edit history, building therefore a collection of N_E text documents, N_E being the total number of edits, so that each document corresponds to the text contained in the page at a certain edit time. We then extracted the 100 most frequent terms (defined as n-grams, sets of ordered words with $2 \leq n \leq 5$ [13]) over the whole collection using the CorText Manager platform (http://www.cortext.fr/projects/cortext-manager), which contains a lexical extraction tool based on a statistical analysis of the n-grams present in a collection of texts and extracts the N most relevant ones, i.e. terms featuring both high *unithood* and high *termhood* as defined in [15]. Time series resulting from this process render the absolute frequency of each term at each version of the Wikipedia article.

To illustrate how the method might work on a historic event from a moderate far past, we analyze a pivotal moment in the American Civil Rights Movement, the Montgomery Bus Boycott. In 1955, Alabama and many other states in the southern region of the United States had Jim Crow laws that enforced racial segregation. These laws prohibited black people from using public facilities that were available to white people. There were separate schools, restaurants and water fountains. White people rode in the front seats of public buses and black people had to sit in the back. Violations could be met with severe repercussions.

One day a black woman, Rosa Parks, decided she would no longer tolerate this unequal treatment. She was "tired of giving in" [18, 17] and deliberately sat in the front seat of a bus. The bus driver told her to move to the back and she refused. Then he called police, who arrested her. As Parks sat in jail, her community rallied and formed the historic boycott. For nearly a year, black

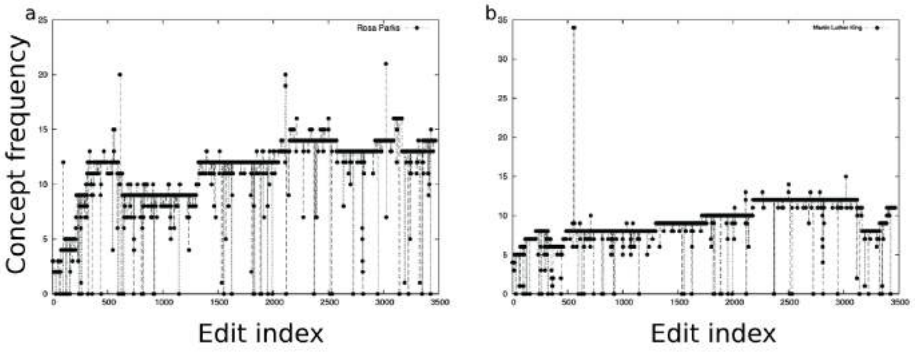

Figure 6: **Prominent characters in the story. a** Frequency of the n-gram 'Rosa Parks' over the different edits of the investigated wiki site. The name of the most prominent figure in the narrative occurs often and fairly regularly, as might be expected. **b** Occurrences of the n-gram 'Martin Luther King'. A more widely recognized hero than Rosa Parks, Martin Luther King was still a young man during the Montgomery Bus Boycott, only beginning to emerge as a Civil Rights leader. Here he appears with about the same frequency as Rosa Parks. Both graphs display a spike around timestamp 600, suggesting a comparison of the two leaders in that edit.

people stayed off the public buses in Montgomery. They walked, organized carpools and created their own bus service. This boycott was one of the sparks of the Civil Rights Movement, which resulted in eradication of the Jim Crow laws and mandates for racial integration of American society.

The associated Wikipedia page [23] details this narrative through 3500 page edits. Our semantic analysis of the page yields the following results shown in figures 6, 7, 8, and 9. Although the effect is notably weak, and we are very far from claiming any finding, the most interesting chart for us would be figure **7a**. There we see the absolute frequency of the n-gram 'Jim Crow Law' as the article evolves towards its current stage. Unluckily, the frequency of this term is very low in general, but the dynamics where a concept rises as the text matures would be definitively interesting for our purposes.

While this basic semantic analysis does not achieve any automated understanding of the narrative, the analysis is useful in assisting comparisons and raising questions and hypotheses to guide closer reading of the text. The method would have been useful in creating an educational program developed nearly 20 years ago, several years prior to the beginning of Wikipedia. Tired of Giving In [5, 21] was an interactive narrative through which viewers could learn about the history of the Montgomery Bus Boycott by hearing moments of the event recounted from the perspectives of different people who participated in the movement or who observed or studied it. Images from Civil Rights archives supplemented the spoken texts in response to viewers' queries. The developers of this educational program spent months collecting and combing through

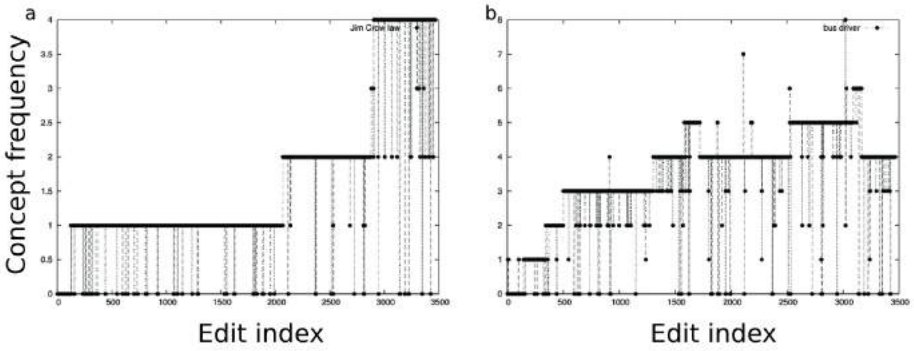

Figure 7: **Concepts that provide context. a** Occurrences of the term 'Jim Crow Law'. *Montgomery Bus Boycott* reaches over time its actual stage. The n-gram 'Jim Crow Law' appears early in the narrative, framing the bus boycott in terms of its social and legal context. It recurs with increasing frequency as the page matures, suggesting the editors interest in both cause and effect of the bus boycott. **b** Occurrences of the term 'bus driver'. The term 'bus driver' also occurs throughout, but less often than 'Rosa Parks'. The spikes around timestamps 2200 and 3100 match those in the 'Rosa Parks' graph, suggesting that those edits may include descriptions of their altercation.

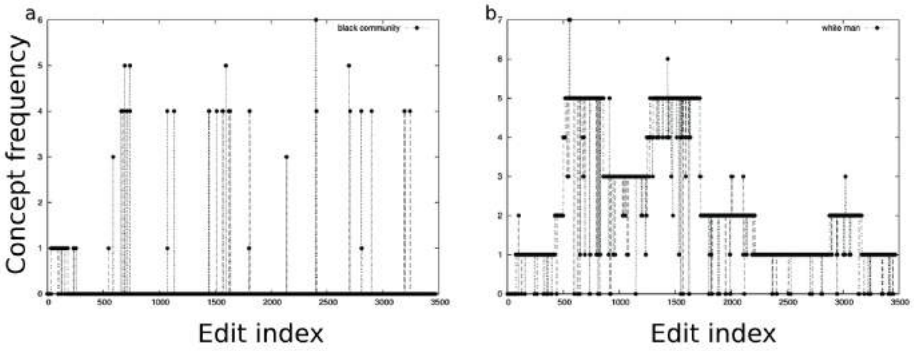

Figure 8: **'Black community' vs. 'white man'. a** Occurrences of 'black community'. A comparison of this term with 'white man' **b** suggests an ironically high relative occurrence of the latter. Overlaps around timestamp 1500 may be particularly interesting to explore.

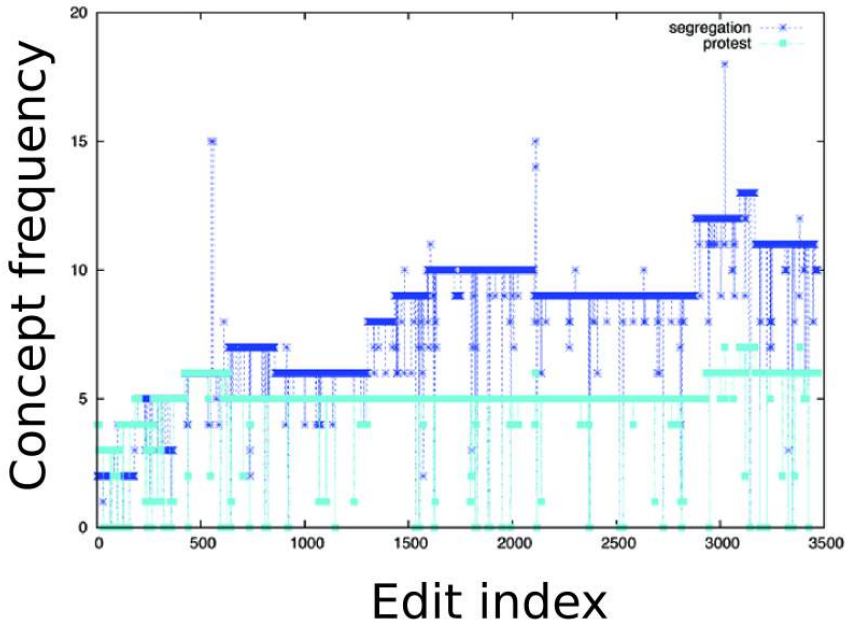

Figure 9: **Occurrences of the terms 'segregation' and 'protest'.** This chart indicates the problematic condition of segregation setting the stage for the protest. The two keywords occur throughout, with the theme of segregation prevailing in the discussion.

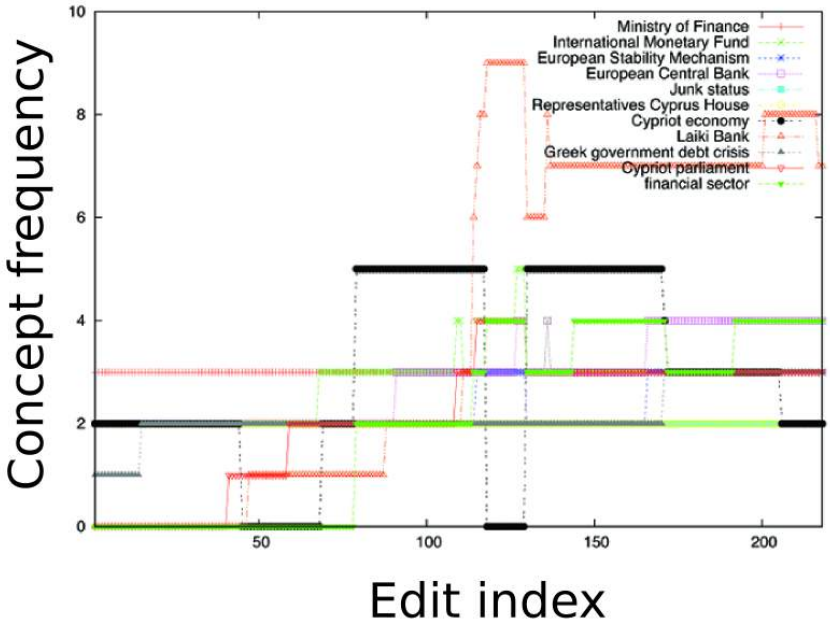

Figure 10: **Occurrences of different keywords in the Wikipedia article for the 2013 Cypriot financial crisis [24].** We can see the rise of a cluster of concepts as the crisis unfolds. Especially the Laiki bank (also known as Cyprus Popular Bank) gains a notoriety without precedent as it is forced to stop money withdrawals from ATMs.

recorded texts and images in order to select information for presentation via the interactive format. Semantic analysis of Wikipedia data could have served as an initial step saving time and effort, and perhaps improving the end result.

In Figure 10 it is illustrated the application of the described methodology to an ongoing historical process. The Cypriot financial crisis [24] took place – and perhaps is still taking place at the time of writing this document – during the first half of 2013 and can be fairly considered one more chapter of the largest financial crisis taking place in the eurozone at the time [25]. In figure 10 we see how different concepts vary in importance over time. Now this can be hardly attributed to a maturing of the text alone, since the crisis was unveiling at the same time as the site was being written. It will be interesting in future work to look at other examples that took place over the last ten years. We focused on this small example because of technical constrains, since the method requires downloading and storing all the versions of a wiki site independently.

4 Conclusions and future work

In this report we propose the study of written history as a complex system. We provided examples of interesting dynamics that can be observed in written history and suggested metaphors with elements from ecology and biology that bring to mind interesting toolsets. An important indicator of how complex written history is—and thus of how worthy it might be for our research community to study—is the lack of results from our methods. While we can safely apply our tools in some data sets, it is an extremely complicated task to solve any of the proposed questions or to link any of our significant outcomes to the phenomenons that we wish to study. In that sense, this report should be read as a reflection about the convenience of using these tools, how to use them, and on what data.

Regarding our datasets, the material we analyze should be the fabric that constitutes written history itself, and until recently this was found solely in books and journals. We decided to make use of Wikipedia, which offers abundant data that is readily accessible and easy to parse. We believe that the phenomenons we aim to study and understand affect this encyclopedia; but we acknowledge that its growth is mainly driven by other, more important forces, and that it is very difficult to tell these factors apart, as usually happens in complex systems. In short, Wikipedia's internal dynamics are likely to shadow written history with the tools we employed. On the other hand, some of the aspects that we wished to investigate require records over a long stretch of time. These would be phenomena associated to generational shifts or to a change of paradigm in the societal conjuncture, which does not happen every day. In that sense, Wikipedia would not be a good data set to work with, since its existence dates back to only a decade ago.

Regarding the methodology, during this summer we investigated two different approaches, one based on networks of correlated concepts and another one based on the semantic study of the evolution of Wikipedia articles over time. A third method has been loosely outlined 1.1 with which we intend to derive rigorous information-theoretical measures not only of Wikipedia pages, but of written accounts of historic events in general.

In section 2 we produced networks of concepts linked by some seed historic event. Taking the temporal series of edits of the pages selected, we worked out the correlation matrix and compared it to a null model. Regarding this null model, we can argue that there is a significant increase in the correlation for the whole of the network. That is to say, one by one we cannot claim that the editing histories of any two concepts are significantly correlated, but taking all the pages as a whole, we appreciate a largely significant increase of the correlation with respect to our reference for neutrality. We would like to give reasons for the observed correlation, but our tools do not allow us to do that yet. The causes for this observed increase in correlation could stem from sociological facts, for internal rules of written history in the search for consistency – as we wish to (but cannot yet) argue, etc. Future work should deepen the potentials for ascertaining the reasons for the correlations we observe. Of course, the null

model should also be considered with care, as we argued in section 2.1. The difficulty of building a null model for the studied datasets should also be noted.

In section 3 we exposed an approach to written history based on the semantics of the texts. We used natural language processing tools and careful, parsimonious data-mining work that requires further human attention. (Indeed, an improvement of the computer tools for this task is another line of research that begs development.) By doing this, we can extract important concepts, characters, and relationships for understanding the unfolding of a historic event. There are two important contributions of this line of thinking:

- Analyzing changes to characters and themes in the written history of a previously recorded historic event, we can guess what the societal zeitgeist considers important in different eras for the understanding of a historic process that is already concluded and that should not, in principle, change anymore. We hypothesize that changes in *the way that society understands* a historic event should follow definite rules that could be guessed from observing the trends that we derive with this methodology.

- For an ongoing event, we expect that the trends and characters involved in it vary much more than for past processes. The main driving force here should be the development of the historic incidents themselves, whose dynamics might be tangentially (or profoundly) grasped by the written history dynamics. In this case, the semantics line of research would provide us with a first coarse graining of complex processes that might be interesting to study. For this report we followed the very recent 2013 Cypriot financial crisis and we could observe how important actors in that stage have risen and fallen in the course of just months.

Altogether, we confirm that written history is an extremely complex system and we believe that the investigation of written history should be considered and studied in this way, and that it will take serious and lasting efforts before any result can be derived. At this point, we are far from suggesting or devising comprehensive theories about written history (theories that should, in any case, be intertwined with those pertaining to other socio-political and cultural processes). Therefore, we propose that lines of research in the immediate future should focus on observing data and trends, and accessing corpora of historic documents that span much larger time scales. We suggest that this can be possible in the mid-term future, thanks to the recent development of relevant tools [13].

As for the methodology reported in this text, we think that the semantic approach to written history is unavoidable and promising, and that it can be useful in providing data that is easier for a researcher to interpret (as suggested above, through an initial coarse graining of a huge flow of information). Meanwhile, any analysis based on correlation networks will require hard work before any conclusion can be drawn about processes of written history. It has been interesting to give it a chance and devote a thorough reflexion to it, and during that process we have learned valuable lessons about correlation and statistics,

but by now it does not seem like a tool ready to be used. Finally, we think that metrics based on information theory as those proposed in [7, 8, 20, 26] would be relevant and would enable quantifying and discerning specific issues in our complex system.

A final thought linked to the excessive task that we bumped against is the following: we believe that much would be gained if the study of written history would have been divided into smaller parts, if we would have focused on a very specific question that we could readily answer with any available tools. Instead of that, we feel that we have undertaken the project from a very general point of view. Examples of such smaller questions could be, to take one single historic topic, however small it may be, and research specific aspects of the coarse graining provided by the semantic analysis. One example of such a specific aspect could be how the number of key concepts evolves over time, a very particular and accessible question.

Acknowledgements

This has been a pretty intense summer and it deserves an acknowledgements section: one that is usually not appropriate to write and one, that is complex enough to never be properly written. We acknowledge the whole and the parts, because there is not one without the other. We acknowledge the generations before, which we could sense breathing information through us (and) into the future. We acknowledge Santa Fe and the Santa Fe Institute, now and forever sacred. We acknowledge the desert, the drought, the sky so blue and crackling with fire, the coyote and the rattle snake, the black bear and the road runner; we acknowledge the trees and the moon, the closest to the earth in years.

And we must write down these names full of meaning: We thank Sander Bais, Tom Carter, John Paul Gonzales, and Juniper Lovato (in strict alphabetical order) for making the impossible possible. We thank Nix Barnett for his priceless helping hands. We thank all the brilliant lecturers we had and the indispensable— and equally brilliant—crew at the Santa Fe Institute. And I am sure that as we finish this acknowledgements sections we are thankful that these last months have not actually been a dream.

References

[1] B. A. Pendleton A. Kittur, B. Suh and E. H. Chi. He says, she says: Conflict and coordination in wikipedia. In *25th Annual ACM Conference on Human Factors in Computing Systems*, 2007.

[2] B. A. Pendleton B. Suh A. Kittur, E. H. Chi and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Alt.CHI at CHI*, 2007.

[3] P. Aragón, A. Kaltenbrunnen, D. Laniado, and Y. Volkovich. Biographical social networks on wikipedia – a cross-cultural study of links that made history. http://arxiv.org/abs/1204.3799.

[4] W. Burkert. *Creation of the Sacred: Tracks of Biology in Early Religions.* Harvard University Press, Cambridge, Massachusetts, 1996.

[5] L. Friedlander C. Strohecker, K. M. Brooks. Experiments with the theatrical greek chorus as a model for interactions with computational narrative systems. In *Narrative Intelligence: Advances in Consciousness Research*, John Benjamins, Amsterdam.

[6] G. J. Chaitin. On the length of programs for computing finite binary sequences. *Journal of the ACM*, 13:547–569, 1966.

[7] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8:17–24, 2012.

[8] Jim Crutchfield. Lectures on "information theory and intrinsic computation" in 2013 complex systems summer school at santa fe.

[9] B. Huberman D. Wilkinson. Cooperation and quality in wikipedia. In *WikiSym '07*, Montreal, Quebec, Canada, 2007. ACM.

[10] N. Chomsky E. S. Herman. *Manufacturing Consent.* Pantheon Books.

[11] J. Kriss F. B. Viegas, M. Wattenberg and F. van Ham. Talk before you talk: Coordination in wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, 2007.

[12] S.A. Frank. Natural selection maximizes fisher information. *Journal of Evolutionary Biology*, 22:231–244, 2009.

[13] A. P. Aiden A. Veres M. K. Gray The Google Books Team J. P. Pickett D. Hoiberg D. Clancy P. Norvig J. Orwant S. Pinker M. A. Nowak J-B Michel, Y. K. Shen and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182, 2011.

[14] E. Szathmry J. M. Smith. *The Major Transitions in Evolution.* Harvard University Press, 1995.

[15] B. Umino K. Kageura. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289, 1996.

[16] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Infromation Transmission*, 1(1):1–7, 1965.

[17] G. J. Reed R. Parks. *Quiet Strength.* Zondervan Publishing, 1994.

[18] J. Haskins R. Parks. *Rosa Parks: My Story.* Dial Books, New York, 1992.

[19] C. Shalizi. Advanced data analysis from an elementary point of view. Preprint of book found at http://www.stat.cmu.edu/∼cshalizi/ADAfaEPoV/.

[20] C. Shalizi. Complexity, entropy and physics of gzip. http://vserver1.cscs.lsa.umich.edu/∼crshalizi/notabene/cep-gzip.

[21] C. Strohecker. See http://www.carolstrohecker.info/projectpages/togi.html for additional publications explicating the 'tired of giving in' project.

[22] J. Voss. Measuring wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics*, 2005.

[23] Montgomery Bus Boycot: http://en.Wikipedia.org/wiki/Montgomery_Bus_Boycott, 2013

[24] 2012-13 Cypriot financial crisis: http://en.wikipedia.org/wiki/2012-13_Cypriot_financial_crisis

[25] Eurozone: http://en.Wikipedia.org/wiki/Eurozone_crisis, 2013

[26] Personal communication with N. Barnett.

[27] Atlantis: http://en.wikipedia.org/wiki/atlantis, 2013.