A Numerologic Analysis System:

Cabbalistic Conclusions from Empirical Data-Mining
A Proposal for the Turbulence.org Web Site
19 January 2001

By Golan Levin (golan@flong.com), with:
  Martin Wattenberg
  Shelly Wynecoop
  David Elashoff
  David Becker

Contents
  Abstract
  Introduction
  Deliverables
  Contributor Biographies

Abstract
The artists conducted an exhaustive empirical study, with the aid of custom software, public search engines and powerful statistical techniques, in order to determine the relative popularity of every integer between 1 and one million. The resulting popularity profile exhibits an extraordinary variety of patterns and singularities which reflect and refract our culture, our minds, and our bodies. For example, certain numbers, such as 212, 486, 718, 911, 1040, 1492, 1776, 68040, or 90210, are observed to be more “popular” than their neighbors because they happen to be used to denominate area codes, tax forms, computer chips, famous dates, or television programs that figure prominently in our culture. Meanwhile, regular periodicities in the histogram, located at multiples of 10, 50, 100 and 1000, clearly mirror our tendency to prefer or create round numbers, while certain numbers, such as 12345 or 8888, appear to be more popular only because they are easier to remember.

We surmise that our dataset is a numeric snapshot of our collective unconscious, and propose herein to return our analyses to the public in the form of an interactive visualization. The visualization will be implemented as a Java applet, to be hosted on the Turbulence.org web site, that permits users to browse our data in such a way as to support both goal-directed exploration as well as undirected foraging. Ultimately, our aim is to provoke our interactants into a consideration of their own numeric manifestations, and to encourage them to draw their own conclusions about our culture's.
Introduction

Humanity’s fascination with numbers is ancient and complex. The history of written language reveals that most of the oldest documents were made exclusively for bookkeeping purposes, and contain only numbers in their enumerations of sheep or lunar cycles. Our present relationship with numbers reveals both a highly developed tool and a highly developed user, working together to measure, create, and predict both ourselves and the world around us. But like every symbiotic couple, the tool we would like to believe is separate from us (and thus objective) is actually an intricate reflection of our thoughts, interests, and capabilities. One intriguing result of this symbiosis is that the numeric system we use to describe patterns, is actually used in a patterned fashion to describe.

Setting the tremendous epistemological implications aside, this interaction can be found with every tool we use: we do not employ a tool without engaging biases about its usefulness. For example, our patterns of hammer-use depend greatly upon what we consider to be “hammerable.” As it turns out, our descriptive application of the number system contains the same sorts of biases. When we choose to say there are “a lot” of gorillas, or “17 gorillas,” or about “20,” we are making a choice based on our interests and abilities. And those interests and abilities are systematically expressed in the numbers we choose to send out into the world.

The Internet provides us with a large and diverse database of the stuff people choose to send out into the world. Among the stuff people put out are large numbers of numbers. By doing a massive automated Internet search on each of the integers from 1 to 1,000,000 and counting the hits for each, we have obtained a picture of the Internet community’s numeric interests and abilities. (An investigation of this nature would have been onerous or impossible before the Internet.) Similar searches executed several months later offer a time-lapse comparison.
The histogram on the previous page shows the popularity of the numbers from 1-2500 collected from the Altavista search engine. Notice the general trend for the lowest integers to have the most hits; this reflects the fact that easily-countable numbers (and especially the numbers countable only on our hands and feet) are more “common” in our lives than larger numbers. Integer multiples of 5, 10, 50, and 100 are visibly emphasized as periodic spikes. A “plateau” rapidly drops off after the 3-digit numbers end; this could be a result of the preponderance of phone numbers and other numeric systems limited to 3-digit clusters. Marked prominences occur between the numbers 93-99 and 1650-2000; this can be attributed to the special use of those integers to represent the year, and in particular the years since the World Wide Web’s growth began to boom. By examining the data more closely, we can observe an over-representation of items illustrating commerce (386, 486, 1040, 68040), computer science (64, 128, 256, 512), the numeric apparatus of human physiology (powers of 10), American culture (666, 911, 90210), history (1492), and many other snapshots of our numerical lives. A yet more discerning eye can pull out items related to numeracy, cultural dominance, and cognitive structure.

By viewing the image differently, wholly different patterns can be brought into focus. In the image below, for example (prepared by Martin Wattenberg), each row represents 100 numbers; the darkened diagonal stripe at the lower left makes visible the special prominence of reduplicative numbers such as 1515 or 2323. Why these numbers are more popular is anyone’s guess; that they are more popular offers fascinating food for thought.

While some patterns are readily evident—such as the tractability of small numbers and the powers of 10—much of the value of this incredibly rich data set must be sought out by an inquisitive interactant. It is in that spirit that we would like to make it available for testing others’ theories about our culture’s use, abuse, or simple capacity for numbers. In the next section, we detail our proposed deliverable, which will allow people to explore the data for themselves.

The numeric system has helped document the regularity and periodicity inherent in our environments and ourselves for centuries. We hope this data will not only shed some light on previously undiscovered regularities concerning our numeric capacities, but also suggest the influence technology has had in changing the set of numbers we can and cannot imagine.
Deliverables
The project members will deliver an interactive Java applet, to be hosted on the Turbulence.org web site. This applet will allow users to browse and explore the dataset we collected. A rough and approximate sketch of the Java applet is shown in the image below.

The applet uses several complete datasets, collected on several different occasions from several different search engines. Special statistical techniques were used to model and analyze the data, and to reduce under- or over-representation errors due to the presence or absence of comma separations in the original source material.

The applet will visualize the data in several different ways simultaneously, all of which will be dynamically linked together in real time. For example, the applet will provide an interactively-zoomable, linear histogram of the data, at the same time that it exposes other patterns with a grid-based density diagram. For each number, the applet will provide its measured popularity, its predicted popularity (according to a model called the Benford distribution), and the statistical likelihood (according to a Chi-squared test) that a given item’s under- or over-representation could be due to chance alone. It will also provide a series of keywords which, according to our computations, appear to be associated with each number; in the image below, for example, the phrase “toll-free” is seen to be associated with the number 888 (one of the area codes reserved for toll-free calls).

In addition to allowing the user to scroll and zoom through hundreds of thousands of datapoints, the applet will also make available a simple interface for highly targeted inquiries and direct searches. This interface is shown at left below, along with other interface elements for reporting the significance of the results, and customizing the configuration of the applet’s display.

**Numerologic Analysis System**
Finally, the applet will contain a number of other modes which allow people to compile and follow up on their searches. The delivery of the applet's sub-site will also include an essay, with an annotated bibliography, describing some of the notable patterns we have observed.

Contributor Biographies

Golan Levin [Project Director/Lead Designer/Lead Engineer] is an artist and engineer interested in creating experiences that probe supple new modes of expression and revelation. Golan received graduate and undergraduate degrees from the Aesthetics and Computation Group at the MIT Media Laboratory, where his work focused on audiovisual performance systems; prior to this, he worked as a research scientist, interaction designer and artist-in-residence at Interval Research Corporation for four years. Golan has exhibited interactive artworks and performances at numerous venues, including Ars Electronica 1997 and 2000, the SIGGRAPH 1996 and 2000 Art Shows, the 1997 International Symposium for Electronic Art, the San Jose Technology Museum of Innovation, and the Berlin Transmediale 2001; his work is also included in the permanent collection of the American Museum of the Moving Image. Most recently, Golan received the 2000 Prix Ars Electronica Award of Distinction for “Scribble,” a live audiovisual composition and performance. Golan resides in New York City.
http://www.media.mit.edu/~golan/

Martin Wattenberg [Consulting Designer/Engineer] is an artist and designer interested in the expressive content latent in pure data. As Director of Interactive Media at SmartMoney.com, Martin developed the highly popular MarketMap visualization tool, for which he received an honorable mention in the 1998 Prix Ars Electronica.
http://www.bewitched.com

Shelly Wynecoop [Consulting Theoretician] has a background which combines research in philosophy, linguistics, cognitive psychology, and statistics. She is currently Assistant Curator of Digital Media at the American Museum of the Moving Image; prior to this, she completed graduate studies in epistemology at the University of Chicago.

David Elashoff [Consulting Statistician] is a Ph.D. candidate in the Statistics department at Stanford University.

David Becker [Consulting Engineer] is a graduate of the MIT Media Laboratory and a software engineer at Amova.com.