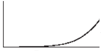
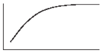
For Good Measure

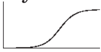
Much Ado about Metadata

DAN GEER AND DAN CONWAY

Dan Geer is the CISO for In-Q-Tel and a security researcher with a quantitative bent. He has a long history with the USENIX Association, including officer positions, program committees, etc. dan@geer.org

Dan Conway is Director of the Center for Business Analytics at Loras College. He has previously served on the faculty at Notre Dame, Indiana University, and Northwestern University. datasciencedan@gmail.com

Trendline metadata inform high frequency trading algorithms, advertising algorithms, and bandwidth balancing algorithms. Trendline metadata inform mitigation choices, budget priorities, and policy. When a trendline measures cumulative events, the shape is called "convex" if the underlying data is increasing in frequency . Similarly, the trendline's shape is called "concave" if the underlying data is decreasing in frequency .

The cumulative life of many kinds of adoption processes take an "s-curve" shape—convex at first and then concave . The s-curve pattern occurs everywhere: it describes the number of VAX computers sold in the 1980s, the prevalence of Internet access in Nigeria, the number of English articles posted on Wikipedia, the spread of cancer, and the adoption of just about any new technology. Parameterized s-curves pinpoint the "inflection point" where acceleration of the growth curve becomes zero and convexity converts to concavity. A parameterized s-curve estimates the final lifetime number of accumulated events as well. (There are lots of references on this if you want to learn more; "logistic" is a good search word with which to begin.)

We were curious about trends in categories of cybersecurity study and whether s-curves might be what we see there. The editors at *IEEE Security & Privacy* very kindly provided us with all keywords from 12 years of articles in *S&P*. Those keywords are certainly varied; there were 7501 keywords—3071 of them unique—spread across 1341 articles. As one might expect, the terms "security" and "privacy" were the most frequently used keywords with 1778 combined occurrences.

The cumulative density of all keywords is shown in Figure 1, indicating a slight increase in the number of keywords used.

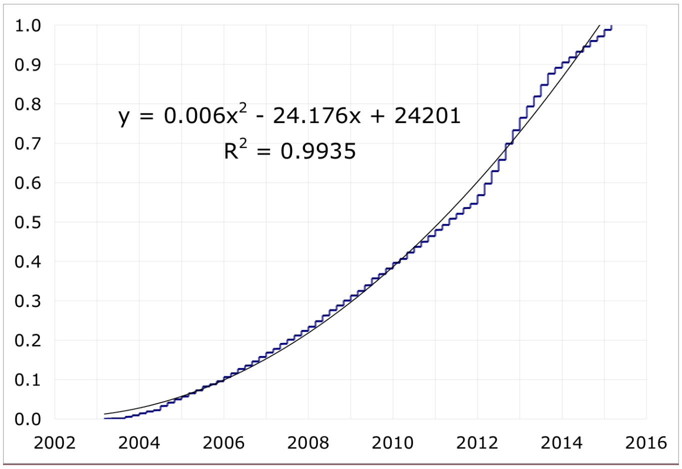


Figure 1: All keywords

Again, a curve that is not only increasing but also increasing at an increasing rate (even if only a slight acceleration) is called "convex." We will use Figure 1 as our basis of comparison to other curves you will see below, and, per convention, we will use twice the coefficient of the x^2 term, 0.006 x 2 = 0.012, as the "convexity" of the curve. The reason we are looking at this is simple: if a given keyword has a convex trendline, then it can be called a "leading indicator" and can be said to be predictive (in context). If the coefficient is negative and therefore the curve's upward growth is decelerating, then it is a "lagging indicator" and can be said to be descriptive.

So how do trendlines in specific keywords compare? If we consider the ten domains of the CISSP Common Body of Knowledge [1], we find a statistically similar pattern to Figure 1 in keyword alignment. The distribution of the domains differs, as "Network Security" appears 10x as often as "Physical Environmental Security," but the overall trendline shapes are identical. They are neither lagging nor leading. We will have to look beyond those domains to find different shapes that tell us more.

Take the keyword "crime"; it has a convex pattern, and Figure 2 tells us that authors' interest in "crime" is increasing (with coefficient .0330).

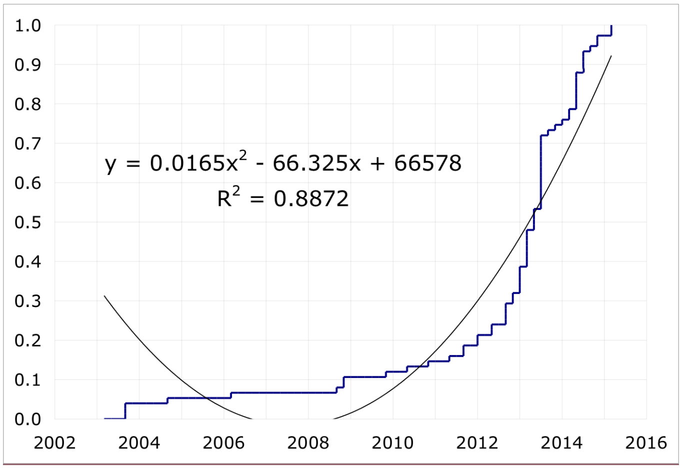


Figure 2: Keyword "crime"

The keyword "honey" (in several merged variations) follows a concave pattern with coefficient –.0017. It is a term that is no longer active in the keyword population, as seen in Figure 3.

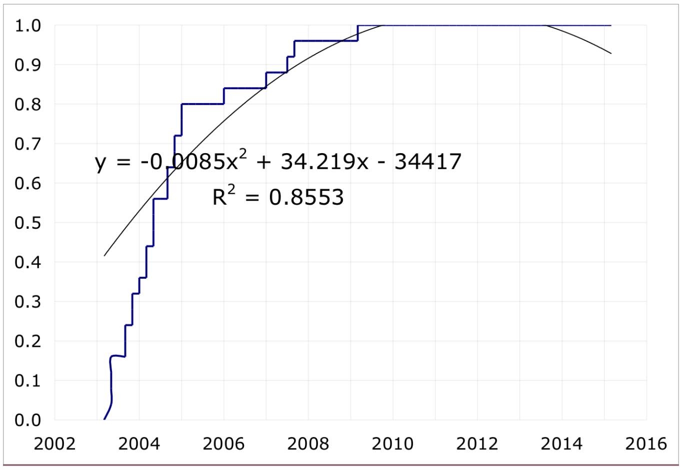


Figure 3: Keyword "honey"

"Virus" (Figure 4) follows a similar pattern to "honey." With a coefficient of –.0072, it has begun to fade away.

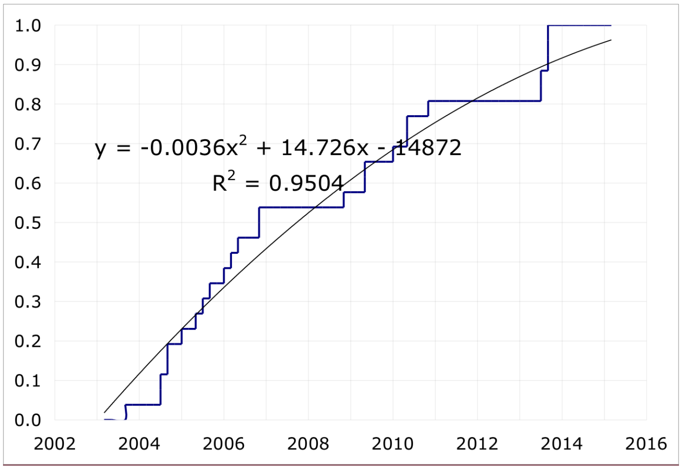


Figure 4: Keyword "virus"

Consider the keyword "identity" as shown in Figure 5. It has been used 56 times in the past 12 years, most recently in March of 2014. This term may be past its peak as indicated by its –.0042 coefficient.

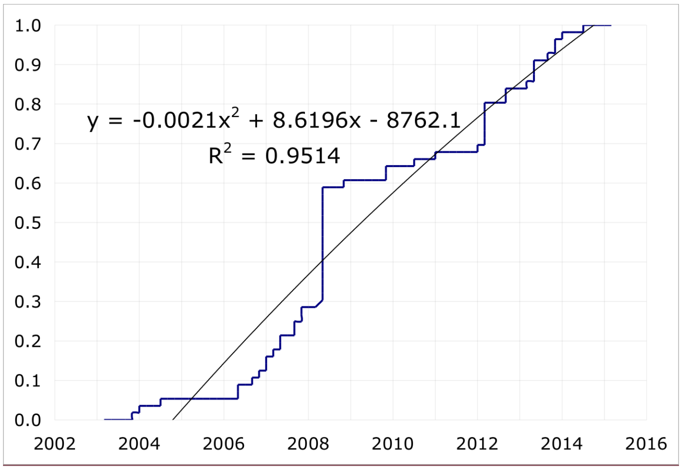


Figure 5: Keyword "identity"

Do these keyword trends reflect actual incidents involving identity? Stipulating that an article marked with the keyword "identity" is not necessarily an article about the misuse of identity information, we nevertheless went to the "dataloss database" [2] and plotted incidents involving loss of any two or more of ACC (account information), ADD (address information), CCN (credit card number), DOB (date of birth), MED (medical information), NAA (name), and SSN (social security number) over the same time period as the *S&P* keyword "identity." For the DatalossDB, we find the nearly linear pattern seen in Figure 6. Not much evidence here of the "identity" keyword being a leading indicator.

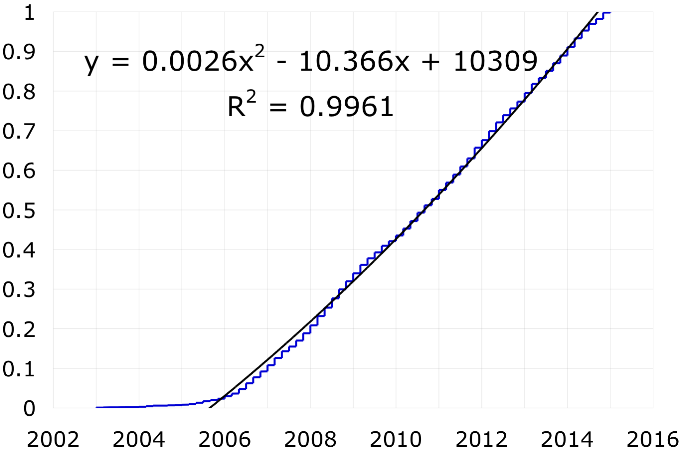


Figure 6: DatalossDB curve for "identity"

(As a side note, Symantec points out [3] that identity theft protection costs approximately $150/year whereas Personally Identifying Information (PII) is available in underground markets for $12–$16 each; hence PII is worth roughly ten times as much to the person identified as to external interests.)

Turning to "crime," we took as our real life measure the incidents reported to DatalossDB as "hack," "stolen [items]," and/or "fraud"—yielding the trendline seen in Figure 7. Compare the convexity coefficient .0186 here for this one measure of actual cybercrime to the .0330 for articles with the keyword "crime" as seen in Figure 2.

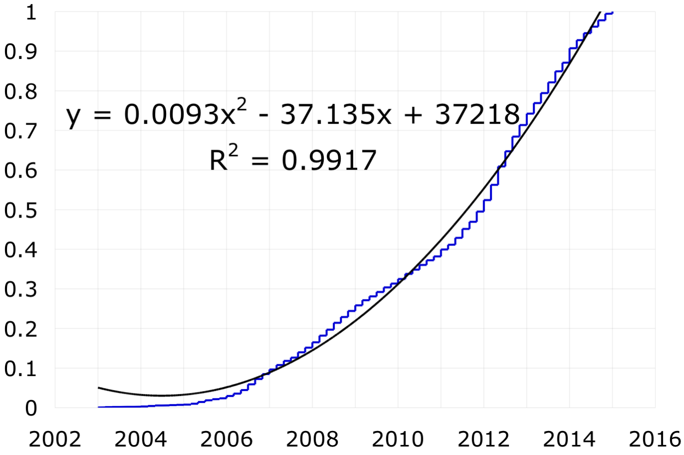


Figure 7: DatalossDB curve for "crime"

In marketing, when a new product is being evaluated, the company will estimate the number and timing of lifetime customers. Customers coming and disappearing soon thereafter are referred to as "churn," and churn models attempt to describe when members will leave a population, be it a population of cell phone customers, those with cancer, or members of an online forum. For example, HP and Cisco measure employee time on LinkedIn and such, attempting to determine who is likely to leave. For those they want to keep, they then intervene. A retailer will want to know the probability of a customer not returning, and send them a coupon to extend the customer lifetime.

If we treat keywords as customers of *IEEE S&P*, then one might ask when "crime" will peak as a customer and begin to decline. This forecasting technique uses the s-curve, which, by definition, is convex up to its "inflection point" and concave after that point. The keyword "crime" appeared 75 times across those 1341 articles. Solving a least squares fit of our s-curve to use of the keyword "crime," we get Figure 8. In other words, we might be now seeing the keyword "crime" start to appear less and less, and we might predict a total of 133 occurrences during its product lifetime (that is to say, before it isn't used anymore at all and the cumulative frequency curve asymptotes).

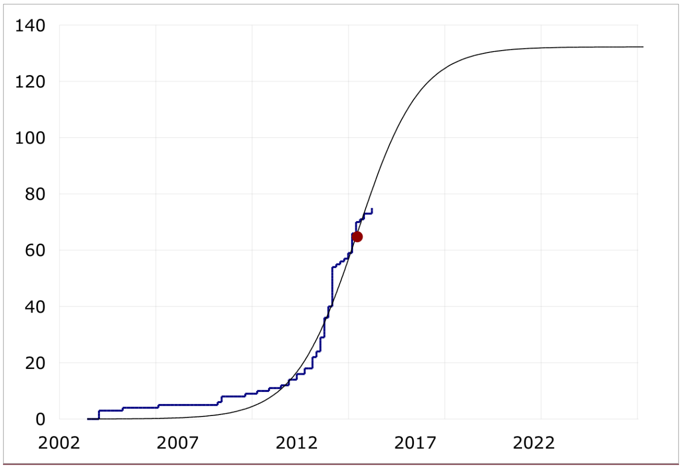


Figure 8: S-curve fit to "crime" keyword, inflection point on July 22, 2014

The term "virus" appears to have hit its inflection point in 2007 and is in active decline. We anticipate very few additional articles on this topic (see Figure 9).

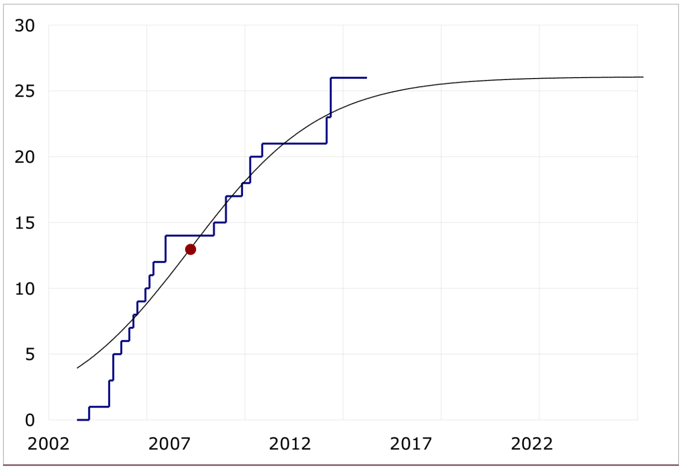


Figure 9: S-curve fit to "virus" keyword, inflection point on November 14, 2007

Let's try a newer term, like "cloud." It certainly came on strong, but is keyword use as shown in Figure 10 telling us that cloud security is approaching a solved problem?

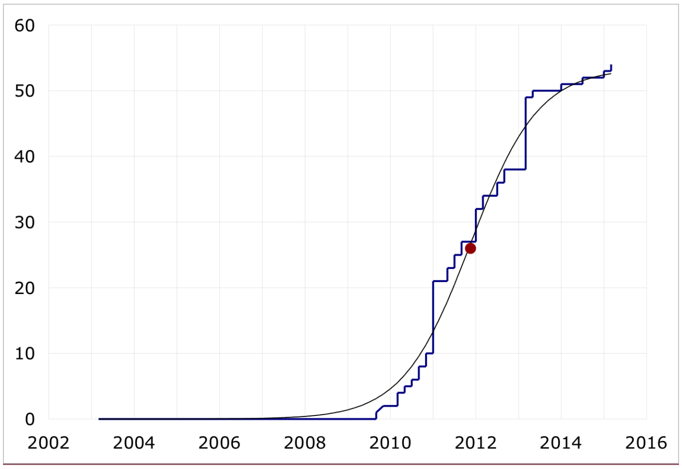


Figure 10: S-curve fit to "cloud" keyword, inflection point on November 17, 2011

Perhaps the keyword "ethic" is your interest, then see Figure 11.

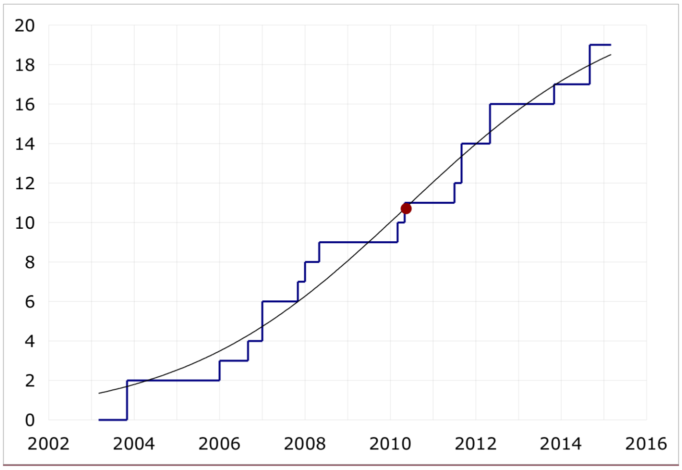


Figure 11: S-curve fit to "ethic" keyword, inflection point on May 16, 2010

The market for solutions seems to follow this lifecycle as well. Virus protection is now available for free, and identity management solutions are decreasing in price. And both terms are past their peak as keywords in *S&P*. Cloud computing adoption rates are clearly not being blocked by security concerns. "Crime" is still trending upward, so can we infer that spending on crime prevention and related services will continue for several years before peaking and beginning its own contraction.

Of course, the keywords chosen by authors of articles are not subject to any particular consistency control. As we said at the outset, there were 3071 distinct keywords across 1341 articles, guaranteeing a lot of singletons (2062 to be precise). We tried binning the keywords, getting far enough to end up with Table 1.

history	8
file types	18
controls	38
meetings	42
security and privacy	54
press	57
targets	71
roles	75
person	135
education	145
metrics	161
countermeasures	171
networks	173
analysis	202
cryptography	255
access control	266
privacy	397
policy	523
attack methods	844
security	1647
<other>	2219

Table 1: Binned keywords

When we binned them, we didn't actually see curves very different from direct use of this or that keyword by itself except for one case: when an author used "security and privacy" as a unitary keyword, rather than "security" and "privacy" as separate keywords, we did get an interesting graph (see Figure 12). Perhaps Figure 12 has something to say about whether "security and privacy" are an indivisible social good or two diverging ones.

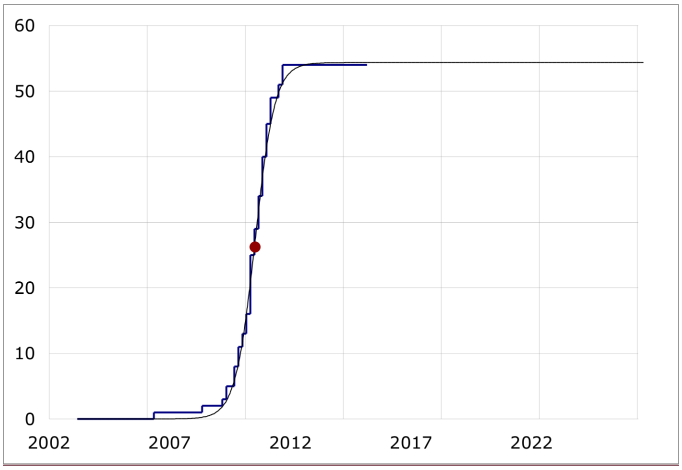


Figure 12: S-curve fit to "security and privacy" keyword, inflection point on July 13, 2010

There's a lot more to explore; we'll be back.

And thanks to you, again, IEEE and DatalossDB.org.

References

[1] https://isc2.org/cissp-domains/default.aspx.

[2] http://datalossdb.org.

[3] http://www.idtheftcenter.org/Privacy-Issues/how-much-is-your-identity-worth-on-the-black-market.html.