The Value of Connecting Things - Part II: The Viking Algorithm

by Mike Bergman - Wednesday, September 03, 2014

http://www.mkbergman.com/1790/the-value-of-connecting-things-part-ii-the-viking-algorithm/

Big Structure Improves Big Data by Orders of Magnitude

Yesterday in the first part of this series we raised the important question of how to value connections made between data. At the Big Data scales represented, we prepared a Basic Facts case of up to 2 billion assertions. We are using asserted "facts" as our value proxy. We'll talk more about value and caveats in the third part of our series, tomorrow.

We saw that early estimates of network effects, such as Metcalfe's law, overestimate value at scale. We looked at Zipf's law as a means to capture the diminishing value of connections given the distance between facts. In today's article we will focus on these factors of interaction and potential value in the specific context of knowledge graphs. Knowledge graphs are Big Structure representations that capture the schematics, concepts and measures in any given knowledge domain (that is, any domain of human activity).

Since I first tried to address the value of knowledge networks some five years ago [1], I have been disturbed about a couple of things [2]. First, I felt that the exponential or geometric bases for estimating the value of information connections were not correct, both because they fail at scale and they don't discriminate that some connections work and are more important, while others are trivial or don't work. Capturing this law of diminishing value in a context that makes sense for knowledge bases was, I felt, the key to answering the value riddle.

I believe we have now, in this series, provided a compelling basis for solving that riddle, which also points the way to further improvements. This assertion is an exciting statement, in that we now may have a quantitative basis in hand for determining where and how to spend our monies for Big Data and Big

Structure initiatives. Such quantitative tools are a huge boon to bring analytic rigor to the data collection and integration challenge.

*Adding connections ("Big Structure") to Big Data can increase the value of enterprise information from **one to three orders of magnitude**; the value also scales linearly with added structure (attributes).*

This article shows that adding connections ("Big Structure") at Big Data scales can increase the value of enterprise information from one (ten) to three (thousands) ***orders of magnitude***. The magnitude of the value scales linearly with each added structure (attribute). These value multipliers from adding Big Structure are a tremendously cost-effective addition to standard data wrangling efforts.

The Value of Knowledge Graphs (VKG) Formulation

The recognition of the need for a law of diminishing returns to reflect the distance between facts or assertions is a central argument in the Briscoe-Odlyzko-Tilly formulation (see [3] directly, and the prior Part I discussion). Not all information is connected, and not all connected information is of equivalent worth. The implied question in these statements, however, is how to capture those differences?

The B-O-T (or sometimes, O-T) formulation does not choose a bad starting proxy for this diminishment law. Zipf's law reflects many observed distributions in human objects, roughly equivalant to power law, Pareto ("80:20") distributions or *n log (n)* diminishing returns with long-tail characteristics. Examples include Internet distributions (such as popularity of Web sites or search terms), human language distributions, income rankings, population distributions, etc. There is no question that Zipf's law distributions are common and frequent.

The only problem with picking the Zipf's law basis, however, is that there was absolutely no evidence that such occurred for information networks or knowledge graphs. Zipf's law distributions tend to be statements across single types for a single attribute distribution. Graphs, we can safely say, are anything but this distribution. Connections and multiple types are the rule, not the exception.

So, maybe the B-O-T formulation was correct, and maybe it was not. There was no empirical evidence to support this assertion for knowledge graphs. And, there did not appear to be a compelling logic argument for relating Zipf's law to graphs other than they are artifacts of human endeavor.

My discomfort in adopting this arbitrary B-O-T basis, even though solidly embedded in human experience, caused me to seek alternative ideas and explanations, but also ones that fulfilled the key structural insights of diminishing returns and non-equivalent assertions that were the focal points of B-O-T, all within a graph context.

The Starting Basis

The breakthough occurred when I discovered an obscure, un-cited paper by Yaakov Stein [4]. Stein, a network and signals processing researcher of the first rank [4], wrote his paper as a means to understand

and quantify his experience of joining LinkedIn and expanding his network. He began without an account and documented his experience as he joined and expanded his network of contacts on LinkedIn. He charted direct links, and then meticulously looked at and recorded secondary and tertiary links.

His formulation recognized that the value to an individual user equaled raising the access to the entire network (*1*) for that user plus the diminishing benefit represented by the participating graph's other participants as measured by average degree of separation (*d*). *d* is an inherent measure of the graph type.

Though his context was a social network, the basic observation obtains: relations diminish by distance within a graph, with average link distance (directly related to degree of separation) being the key relevance metric. Connected "facts" or "friends" is essentially the same thing. It is all about what is shared amongst graph nodes.

The usefulness of this approach is that it grounds the multiplier effect in an inherent characteristic of the source graph, its average degree of separation [5]. Like Zipf's law, the degree of separation is a distance measure, but one grounded specifically in graphs. Here is the Stein formulation:

$$V = n^{(1+1/d)}$$

where V is potential value, n is number of graph nodes, and
d is the graph's average degrees of separation.

A graph with a degree of separation of 4, then, would exhibit a network-wide power factor of 5/4 (4/4 plus 1/4).

The Viking Algorithm

As applied to knowledge graphs, however, this formulation still has two problems. The first minor one is that the degree of separation parameter should be *D* (the average across structure) rather than *d*. The second substantive one is that a correction factor needs to be included that accounts for the probability that an assertion may be false. This factor, *F*, is *1 - the measured error rate*.

The resulting algorithm we term the Value of Knowledge Graph formulation, or the VKG (Viking) algorithm. It is expressed like this:

$$V = F \times n^{(1+1/D)}$$

where V is potential value, F is average assertion accuracy,
n is number of graph nodes, and D is the graph's average degrees of separation.

F is meant to be analogous to F-measure, the combined precision and recall statistic for information retrieval and NLP tasks. *F* in the case of the Viking algorithm is also meant to be a combined statistic that represents the "accuracy" (verifiable truthfulness) of statements asserted in the graph. *F* is essentially an estimated value for the residual falsity for the average statement in a graph, after removal of all assertions

that do not meet existing coherency, consistency or completeness tests. *F* is determined by sampling statements across the graph and manually testing for truthfulness (or in a logical sense, validity given the existing statements in the graph). An *F* of 1 signifies complete truthfulness (accuracy); an F of 0 represents complete falsity [6].

Viking in Relation to Other Network Estimators

Now, with this explanation of basis, we can again look at the value of the Viking (VKG) algorithm in comparison to those discussed in the first part of this series. Again on a logarithmic scale, here are those results:

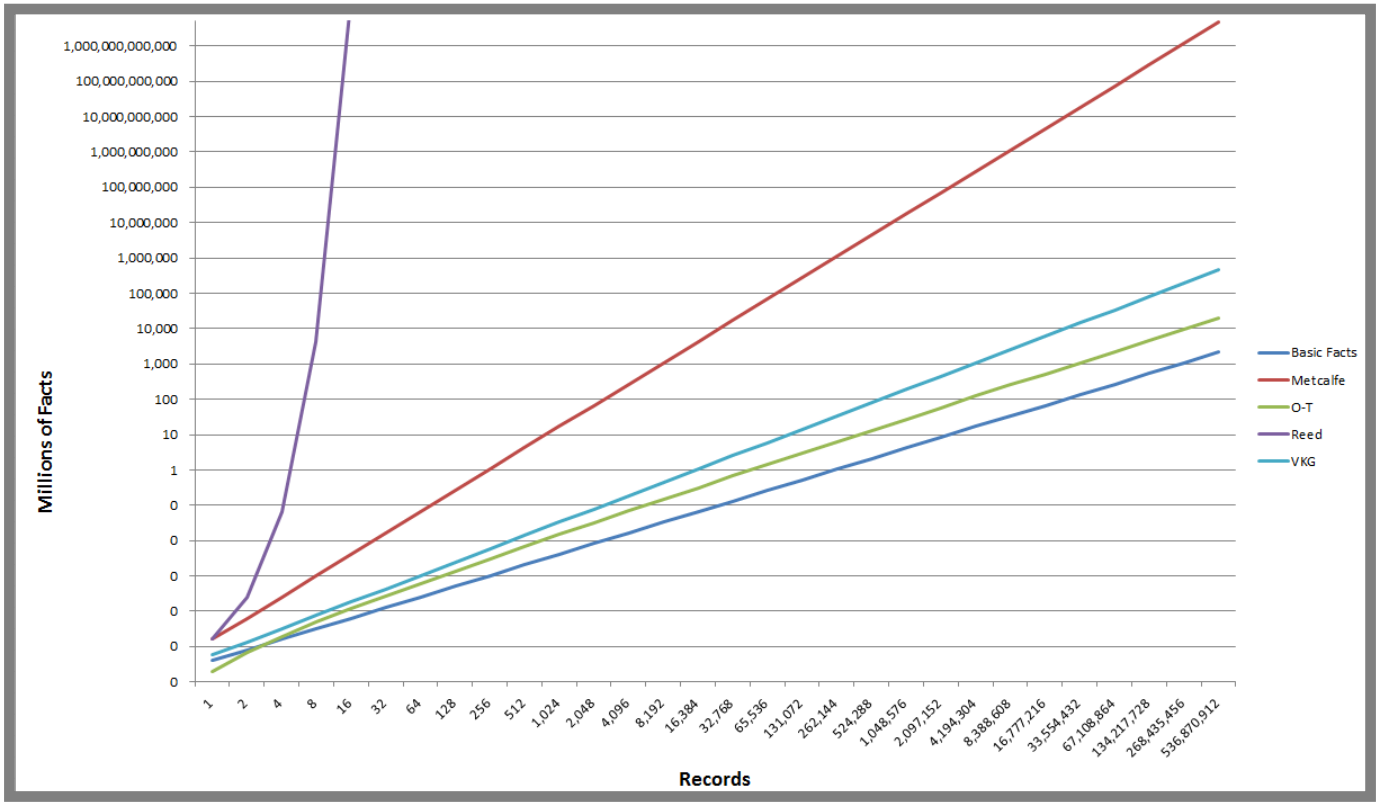

Figure 1. Knowledge Network Estimates

Excluding the exponential and geometric multipliers (namely, the "laws" of Metcalfe and Reed) in the top two curves, this shows the Viking (VKG) algorithm to have higher value than the B-O-T (O-T) algorithm, both of which are considerably higher than the Basic Facts. However, because the **Figure 1** above has a logarithmic scale, these differences are harder to discern.

Viking Benefits Over the Basic Facts

Now corrected with our assumed *F* factor, we can begin to tease out the value benefits of connecting "facts" versus the unconnected Basic Facts. As with any logarithmic function, we see that the value

benefits from connections increase in a growing manner at larger scales. For example, as **Figure 2** shows below, at a level of 1000 records, the benefits from connections are 7x greater than unconnected data. By the time the scale grows to 1 million or 500 million records, the value benefits of connections grows to 44x to 215x, respectively:

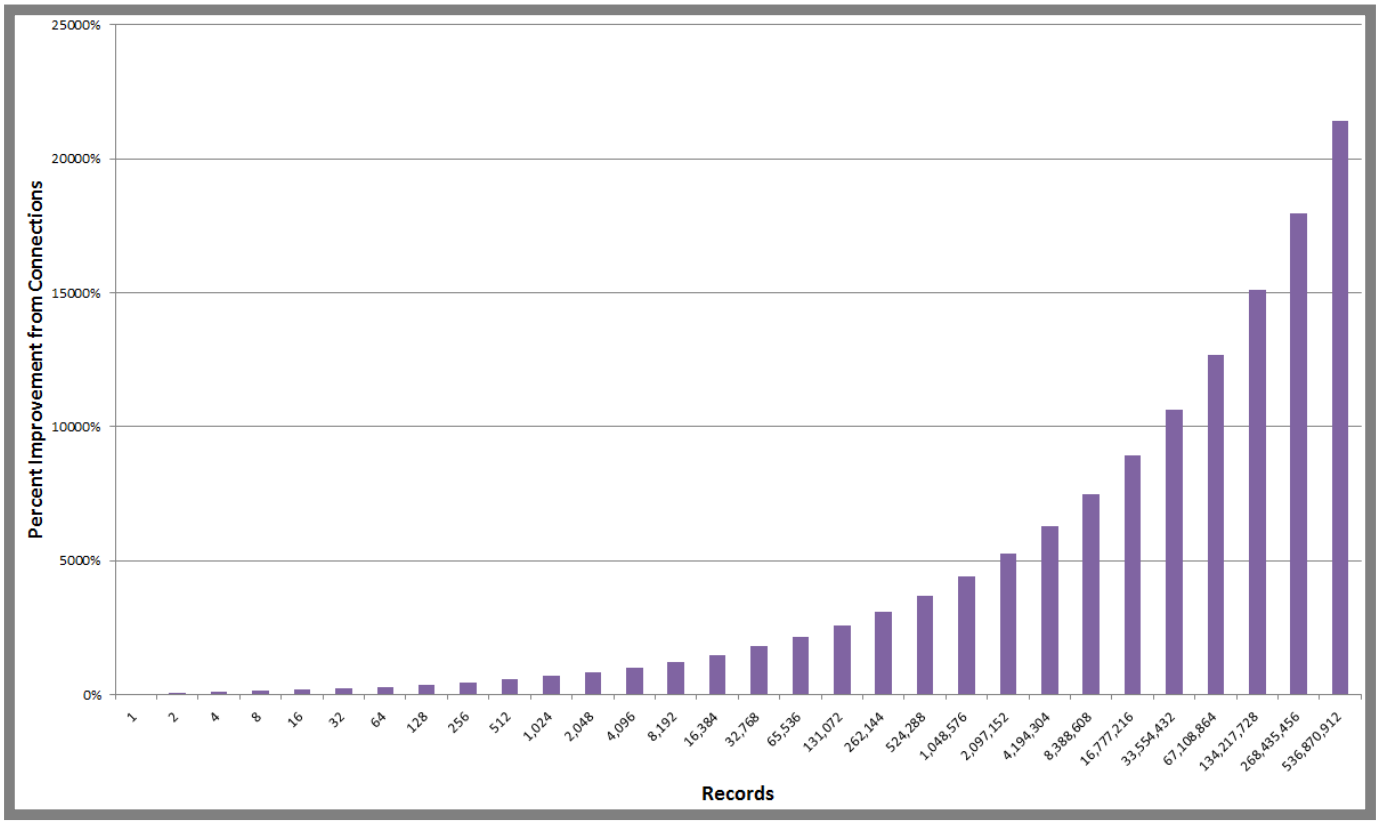

Figure 2. Percent Improvement from Connections Scales with Records Size

Benefits from connections increase as a power function at increasing scales.

Setting the VKG Factor D

But the potential value of connectedness is also a function of the general degree of information separation for the given domain. We are still in the early phases of gathering statistics for such things, but the table below summarizes what is known about the "standard" level of connectivity in various domains and applications. Note, in general, most any knowledge graph would have a *D* factor ranging from 2 to 8:

Category	Degrees of Separation (D)	Notes
Food webs	~ 2	[7]
Genetic differences	~ 3	[8]
LinkedIn	~ 3	[4]
Twitter	3.435 - 4.67	[9]
Facebook	3.74	[10]

Potential research collaborators	~ 4	[11]
UMBEL	~ 5.2	[12]
Social networks (general)	~ 6	
Mobile ad hoc networks	~ 7	[13]
Small-world networks (max)	~ 8	[14]

Table 1. Degrees of Separation for Various Knowledge Networks

More tightly linked, cohesive domains tend to have the lower degrees of separation. It is also interesting to note that some social networks, like Twitter and Facebook, are also able to lower degrees of separation (in comparison to their nominal "social network" benchmark) by virtue of the nature of their service.

As experience is gained and with more research, I expect more estimates and more refined ones. Depending on the nature of the domain at hand, it should then be possible to pick the closest analog to use in the Viking valuation algorithm. Nonetheless, we already have a range and respective values to provide meaningful value estimates today.

Using the values in **Table 1**, we are thus able to plot the effects (again, log scale) of these various degrees of separation in terms of the "fact" assertions that can be made for our Big Data test dataset:

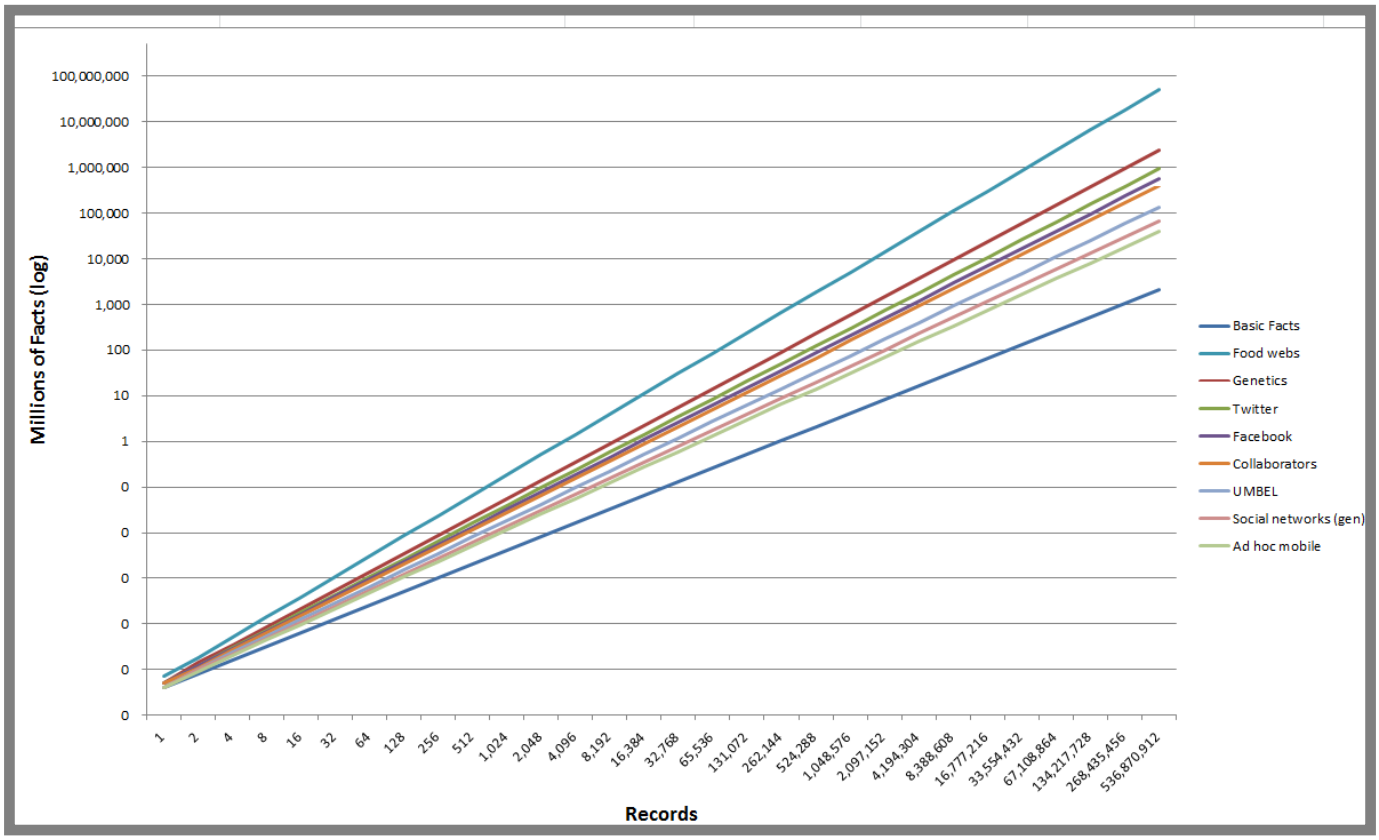

Figure 3. Nature of Knowledge Graph Affects Potential Network Value

At the nominal Big Data scales of 100,000 and 1,000,000 records, the value of data connections in comparison to the unconnected Basic Facts case shows these following value improvement multipliers:

Domain	100,000 Records	1,000,000 Records
Food webs	203x	611x
Genetic differences	38x	84x
Twitter	23x	46x
Facebook	17x	33x
Potential research collaborators	14x	26x
UMBEL	8x	12x
Social networks (general)	5x	8x
Mobile ad hoc networks	3x	5x

Table 2. Multiplier (X) Improvements by Domain from Connections Over the Basic Facts

Of course, our "Big Data Example" from Part I was silent about the exact nature of its knowledge graph. Based on empircial experience to date, the benefits from connecting data that was previously unconnected should fall somewhere within the limits of **Table 2**. Even at rather low scales and more loosely-connected domains, the value improvements in making connections with data is many-fold. At larger scales for tighter networks, the multipliers can become astounding.

Adding Structure to the Underlying Data

Another implication that the Viking algorithm allows us to test is the comparative benefit from adding structure to our datasets. Actually, "adding structure" is not strictly correct; it is "structurizing" the data via characterizations, attributes and categorizations. Of course, not all structure is created equal. Assigning or classifying our records into types, for example, applies to all records across the datasets and provides powerful cross-record linkages. Adding annotations or metadata to single records provides much lower benefits.

When we add structure across datasets the value improvements are a linear percent, as this figure shows:

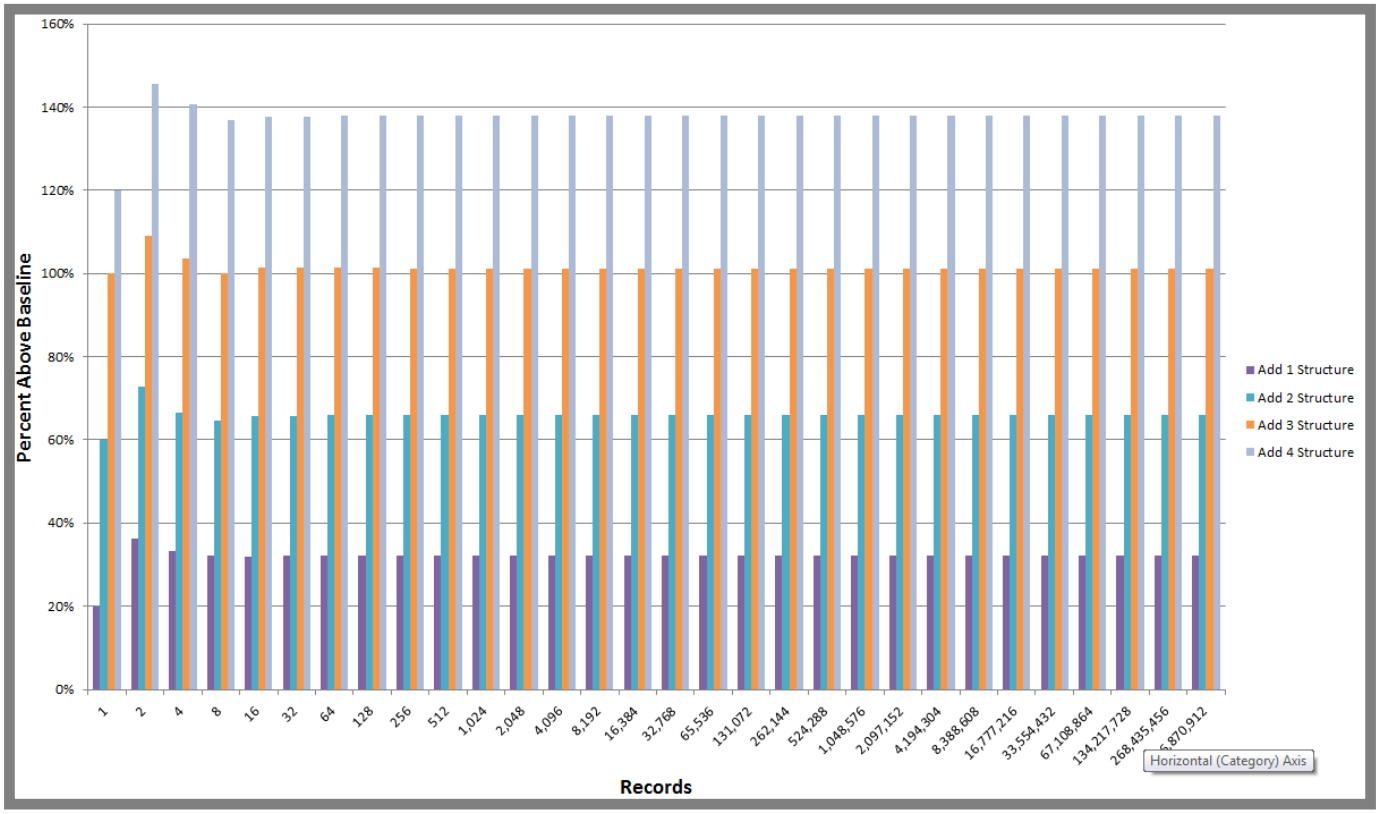

Figure 4. Adding Structure has a Linear Effect on Value

For our Big Data example, each across-dataset structure characterization adds about 25% to 30% value per structure. Adding four structural characterizations, for example, more than doubles the "facts" assertion value (~ 140%) to the datasets.

Preview of Last Part

The last part forthcoming tomorrow will summarize the implications from the Viking algorithm on the role and importance of Big Structure to your organization's Big Data efforts. Some caveats and future directions will conclude the series.

[1] The two articles written at that time were, M.K. Bergman, 2009. Structure the World, in *AI3:::Adaptive Information* blog, August 3, 2009, and M.K. Bergman, 2009, The Law of Linked Data, *AI3:::Adaptive Information* blog, October 11, 2009.

[2] See also the then-current state of analysis by Eric Hellman, 2009. Normal and Inverse Network Effects for Linked Data, published in his *blog*, October 15, 2009; see http://go-to-hellman.blogspot.com/2009/10/normal-and-inverse-network-effects-for.html.

[3] Bob Briscoe, Andrew Odlyzko, and Benjamin Tilly, 2006. "Metcalfe's Law is Wrong," in *IEEE Spectrum*, July 2006. A copy may be viewed at http://www.cse.unr.edu/~yuksem/teaching/nae/reading/2006-briscoe-metcalfes.pdf. Odlyzko, and Tilly had published an earlier version, (sometimes the approach is shown as O-T in addition to B-O-T),

and the basic form of the algorithm appears in a single Odlyzko paper.

[4] Yaakov (J) Stein, 2009. The Value of Being Linked In, on his *personal Web site*, April 2009; see http://www.dspcsp.com/pubs/linkedin.pdf. Note that his empirical tests suggested a degree of separation for LinkedIn of 3.

[5] The average degree of separation is simply the graph's average path distance - 1. For an explanation of average path distance, see [12].

[6] *F* is a summed average value across all assertions within a knowledge graph. In information retrieval, F-measures are now being achieved that exceed 0.90 (90%). For the cases used herein, *F* is estimated at 0.85. Again, this parameter is measured after all standard coherency, consistency, and completeness tests are applied to the ontology. These tests routinely remove many false assertions and establish the basic integrity of the graph. This acceptance threshold is itself constantly improving as experience is gained with basic graph integrity tests. In other words, tomorrow's thresholds will be higher than today's.

[7] Richard J. Williams, Eric L. Berlow, Jennifer A. Dunne, Albert-László Barabási, and Neo D. Martinez, 2002. Two Degrees of Separation in Complex Food Webs, in *Proceedings of the National Academy of Sciences*, 99 (20):12913-12916, September 16, 2002, doi:10.1073/pnas.192448799; see http://www.pnas.org/content/99/20/12913.full.

[8] See http://blog.23andme.com/ancestry/three-genetic-degrees-of-separation/.

[9] Reza Bakhshandeh, Mehdi Samadi, Zohreh Azimifar and Jonathan Schaeffer, 2011. Degrees of Separation in Social Networks, in *Proceedings, The Fourth International Symposium on Combinatorial Search (SoCS-2011)*, 6 pp.; see http://www.aaai.org/ocs/index.php/SOCS/SOCS11/paper/viewFile/4031/4352; and Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, 2010. What is Twitter, a Social Network or a News Media?, in *Proceedings of the 19th International Conference on World Wide Web*, April 26–30, 2010, Raleigh, North Carolina, pp. 591-600, ACM; see http://snap.stanford.edu/class/cs224w-readings/kwak10twitter.pdf.

[10] Lars Backstrom et al., 2012. Four Degrees of Separation, *Archiv.org*, January 6, 2012; see http://arxiv.org/pdf/1111.4570.pdf.

[11] Paweena Chaiwanarom, Ryutaro Ichise, and Chidchano Lursinsap, 2010. Finding Potential Research Collaborators in Four Degrees of Separation, pp. 399-410, in Longbing Cao, Jiang Zhong, and Yong Feng, eds., *Advanced Data Mining and Applications*, Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-17313-4_39.

[12] See http://fgiasson.com/blog/index.php/2014/08/11/graph-analysis-of-a-big-structure-umbel/#Average-Path-Length-Distribution.

[13] Maria Papadopouli and Henning Schulzrinne, 2000. Seven Degrees of Separation in Mobile ad hoc Networks, presented at *Global Telecommunications Conference, 2000 (GLOBECOM'00)*, IEEE. Vol. 3; see http://www.huaxiaspace.net/academic/classes/wi02/cse294/20020222globecom2000.pdf.

[14] Paolo Pin, 2006. Eight Degrees of Separation, in *Nota di Lavoro, Fondazione Eni Enrico Mattei*, No. 78.2006 see http://www.econstor.eu/bitstream/10419/74249/1/NDL2006-078.pdf.
