The Value of Connecting Things - Part I: A Foundation Based on the Network Effect

by Mike Bergman - Tuesday, September 02, 2014

http://www.mkbergman.com/1788/the-value-of-connecting-things-part-i-a-foundation-based-on-the-network-effect/

Teasing Out the Role of Big Structure in Context and Connections

The hackneyed phrase of "connect the dots" reflects our basic intuition that there is value in making connections amongst relevant data. But, what is this value? How might we quantify it? This topic, and a method and guidelines for doing so, are the subject of this article and the second and third parts to follow.

The reason it is important to quantify the value of connected information is that such an estimate helps to define what effort or cost we can justify in order to derive those connections. In Big Data, for example, we already know that 50% to 80% of the costs in assembling relevant datasets is due to data wrangling -- the effort to extract, transform and clean the input data [1]. No where, however, do we know what it is worth to go to the next step of working to connect those data.

Quantifying this understanding will thus also help determine what the value is in developing Big Structure, the approach we have been most recently discussing for how to organize and connect Big Data. Big Structure sets the schematic and data relationships for how data from disparate sources can be connected together.

About five years ago I wrote my first articles on how we might approach the quantification of these information connections [2]. That first, cursory look was useful for bounding the problem, but no firm conclusions as to how to specifically quantify this value were proposed. Like other graphs or networks, the usefulness of the 'network effect' to bound the question was clear. It has taken further research and experiences with actual linked datasets to point to how to resolve this quantification challenge.

Foundations in the Network Effect

The network effect was first realized in the early days of telephone networks, where the value of the system increased as a function of more users [3]. We have also long recognized a similar effect in connecting information together and the breaking down of information or 'data silos'. As the following diagram shows, unconnected data nodes or silos look like random particles caught in the chaos of Brownian motion:

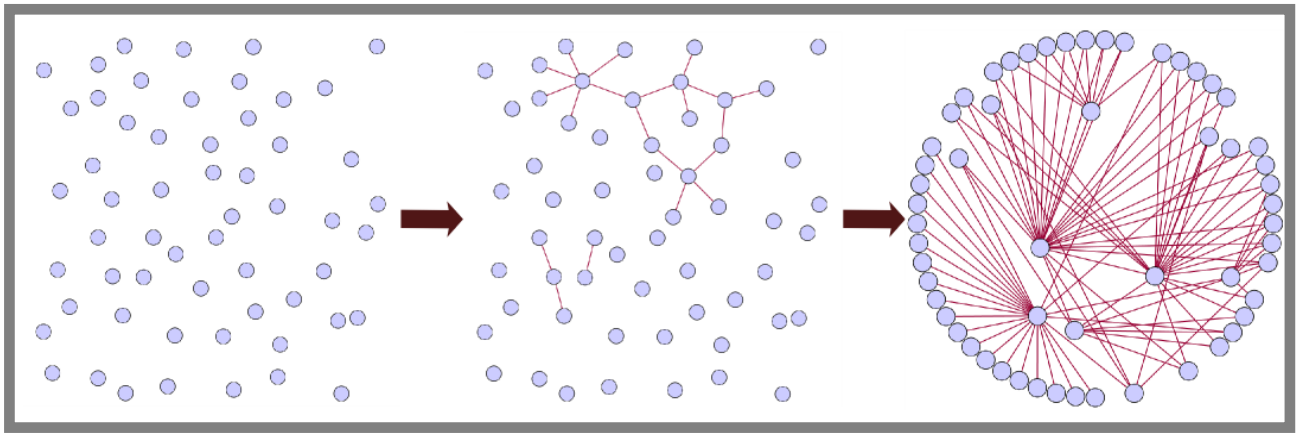

Figure 1. 'Network Effect' for Connected Data

As initial connections get made, bits of structure begin to emerge. But, as connections are proliferated — analogous to the network effects of connected networks — coherence and more structure emerge.

This emergence of structure is particularly evident in physical networks, such as the growth of this hypothetical telecommunications network:

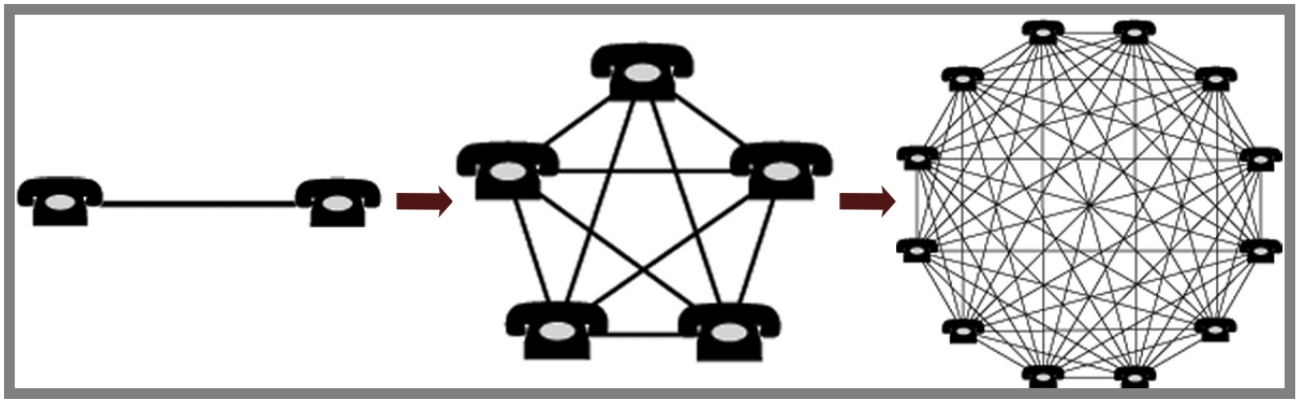

Figure 2. 'Network Effect' for Telecommunications Networks

This diagram, modified from Wikipedia to be a horizontal image, shows how two telephones can make only one connection, five can make 10 connections, and twelve can make 66 connections, etc. It is this

very multiplier effect that has led to most of the thinking of how to quantify the network effect.

We can see an interesting parallel between telecommunications networks and knowledge graphs. In the telecommunication network, the addition of a new user (node) by definition brings with it connections. This is what is shown in **Figure 2**. But in information silos, the information is already there (nodes, or the left-side of **Figure 1**); what is missing are the connections (the right-side of **Figure 1**). By explicitly adding connections we can also create network effects, as others have noted [4].

However, once we understand these parallels, we must also recognize the differences. To properly estimate the network effects of knowledge graphs, we must be explicit about the similarities and differences with other (physical) networks [5].

Objectives for a Knowledge Graph Formulation

Since our objective is to quantify the "value" of a knowledge graph, we must first ask what is the basis of this value. In the best of all worlds, we would know the monetary worth of information, so we could justify what to spend in order to leverage it, which of course varies wildly across bases and sources. But we don't. We do know, however, that a knowledge graph and the information it connects to constitutes a knowledge base. In the context of a knowledge base, the measure of value is the number of "correct" facts it contains. Therefore, we will use the number of connections in the graph (equivalent to the number of triple statements) as a proxy for value, representing the asserted "facts" of the graph.

We will also seek measures of graph distance and connectedness to capture the network-like qualities of the knowledge base. The characteristics of the graph itself should be the input base upon which to estimate value.

Alternative Estimates of the Network Effect

The earliest effort to estimate the value of physical networks was Sarnoff's law, developed by David Sarnoff, for many years the leader of the Radio Corporation of America (RCA). He posited that the value of a broadcast network was directly proportional to its number of viewers (*n*). However, the problem with this formulation is that a broadcast network is only one way, from broadcaster to user. What of networks where there is interaction or two-way linkages? The benefits of such networks must surely be more than linear.

Once we get into interaction effects we get into multipliers. And the proper nature of those multipliers must come from the nature and extent of those interactions, as well as perhaps the nature of the network itself. I discuss below some of the more prominent candidates that have been put forward for estimating the network effect, or the value of networks.

Metcalfe's Law

Metcalfe's law was the first direct derivation from the telecommunications model. Robert Metcalfe formulated it about 1980 in relation to Ethernet and fax machines; the "law" was then named for Metcalfe and popularized by George Gilder in 1993. The actual algorithm proposed by Metcalfe calculates the

number of unique connections in a network with *n* nodes to be $n(n ? 1)/2$, which is proportional to n^2. This makes Metcalfe's law a quadratic growth equation.

The law is generally simplified [6] to state that the value of a telecommunications network is proportional to the square of the number of users of the system (n^2):

$$V = n^2$$

where V is potential value, and n is number of graph nodes.

Gilder's popularization and the early growth of the Internet made the estimation of the benefits of network effects a very timely topic. For example, as a value measure, the network effect could be used to estimate the benefits for larger and larger numbers of users. Some have even blamed Metcalfe's law for contributing to the creation (and then bursting) of the "dot-com bubble" of the late 1990s [7].

Metcalfe's law clearly showed that interaction effects between nodes could generate multipliers that scaled rapidly with increasing numbers of nodes (users).

Reed's Law

From a different perspective and with a different take, David Reed came up with a multiplier formulation that is the largest presented -- anywhere. Reed's context is social groups, and from that perspective he can envision arbitrary sized groups forming amongst any and all participants (nodes). Because of this theoretical, global scope, justified through examples such as eBay and chat rooms, Reed specifically defined group-forming networks (GFNs), as the applicable scope [8]. The simplified formulation for Reed is:

$$V = 2^n$$

where V is potential value, and n is number of nodes.

In scope and context, Reed does not apply to knowledge graphs, and even in the areas of social groups, most researchers find the exponential implications of Reed's law unsupportable [9]. The next group, for example, offers direct criticism.

Briscoe - Odlyzko - Tilly Formulation

Under the provocative title, "Metcalfe's Law is Wrong," Briscoe, Odlyzko, and Tilly challenged both the Metcalfe and Reed approaches in 2006 [10]. Using the proxy of Internet valuation, the authors were able to show how impractical the implications of either approach were at scale. Like the bet of rice (or wheat) doubling each of the 64 squares on a chessboard bankrupting the kingdom, the exponential implications of these two "laws" can be seen to (eventually) violate common sense.

The fundamental fallacy associated with both the Metcalfe and Reed approaches is that all potential links are of equal value [10]. But no where in the real world do we see this to be true. There must be some law

of diminishing returns that must be applied to slow the unsustainable rates of exponential or (to a lesser extent) quadratic growth.

After much hand waving, the authors chose Zipf's law as their basis for this diminishing return. The increasing "decay rate" with distance is a common distribution pattern for real-world datasets, which Zipf's law specifically addresses, always showing power law distributions with long tails. To approximate this distribution they offered the simple *n log (n)* formulation of Zipf's law [11].

$$V = n\ log(n)$$

where V is potential value, and n is number of graph nodes.

This is a reasonable approximation, but one that is never related directly to the nature of graphs or networks. That is the source of the next layer of refinements.

VKG Formulation

I will discuss this algorithm, our recommended formulation, in Part II of this series.

A Big Data Example

In order to discuss further the question of value arising from network effects, we need a case study example. We also need to define "value", which in this case study example, as also noted above, means the number of "facts" or assertions in our database [12]. This we call the Basic Facts ("assertions") column.

For our basic "facts", we consider a data series that is doubling in size for each step, eventually reaching a half billion records. Each record has four attributes or characterizations, leading to a total of more than 2 billion "facts" in the database. These are the first two columns in this table:

Records	Basic Facts ("assertions")	Big Data Example
1	4	4
2	8	8
4	16	16
8	32	32
16	64	64
32	128	128
64	256	256
128	512	512
256	1,024	1,024
512	2,048	2,048
1,024	4,096	4,100
2,048	8,192	8,200
4,096	16,384	16,400
8,192	32,768	32,800

16,384	65,536	65,600
32,768	131,072	131,200
65,536	262,144	262,404
131,072	524,288	524,812
262,144	1,048,576	1,049,624
524,288	2,097,152	2,099,248
1,048,576	4,194,304	4,198,496
2,097,152	8,388,608	8,396,996
4,194,304	16,777,216	16,793,992
8,388,608	33,554,432	33,587,984
16,777,216	67,108,864	67,175,972
33,554,432	134,217,728	134,351,944
67,108,864	268,435,456	268,703,888
134,217,728	536,870,912	537,407,780
268,435,456	1,073,741,824	1,074,815,564
536,870,912	2,147,483,648	2,149,631,128

Table 1. Basic Facts Connections with a Big Data Example

At this point, we have no connections between records. Each record has four attributes each, in isolation. This basis is akin to the unconnected dots on the left side of **Figure 1** above.

For our Big Data Example, we will posit a record matching procedure as our first task for a new Big Data initiative. The assumption is that across all records, one-in-10000 matches another record. This results in the number of assertions ("facts") shown in the third column in the table above. The posited Big Data initiative results in a 0.10% increase in "facts", irrespective of record scale, once the matching threshold is reached. This result is not terribly impressive, but is perhaps not too unrelated from a first foray into a Big Data project.

Note that the following charts and analyses (including in the next part tomorrow) use as their "Basic Facts" the number of "assertions", or the middle column in the table above. Though Big Data may represent an initial 0.10% improvement over this, that is immaterial to what our Big Structure viewpoints will provide. So, our "Basic Facts" will be unconnected records.

Applying Network Effects to the Basic Facts

We can now apply our various network effect estimators to this base case. And, because of the fast-compounding nature of both the Reed and Metcalfe approaches, we need to plot this out on logarithmic scale [13] (click to enlarge):

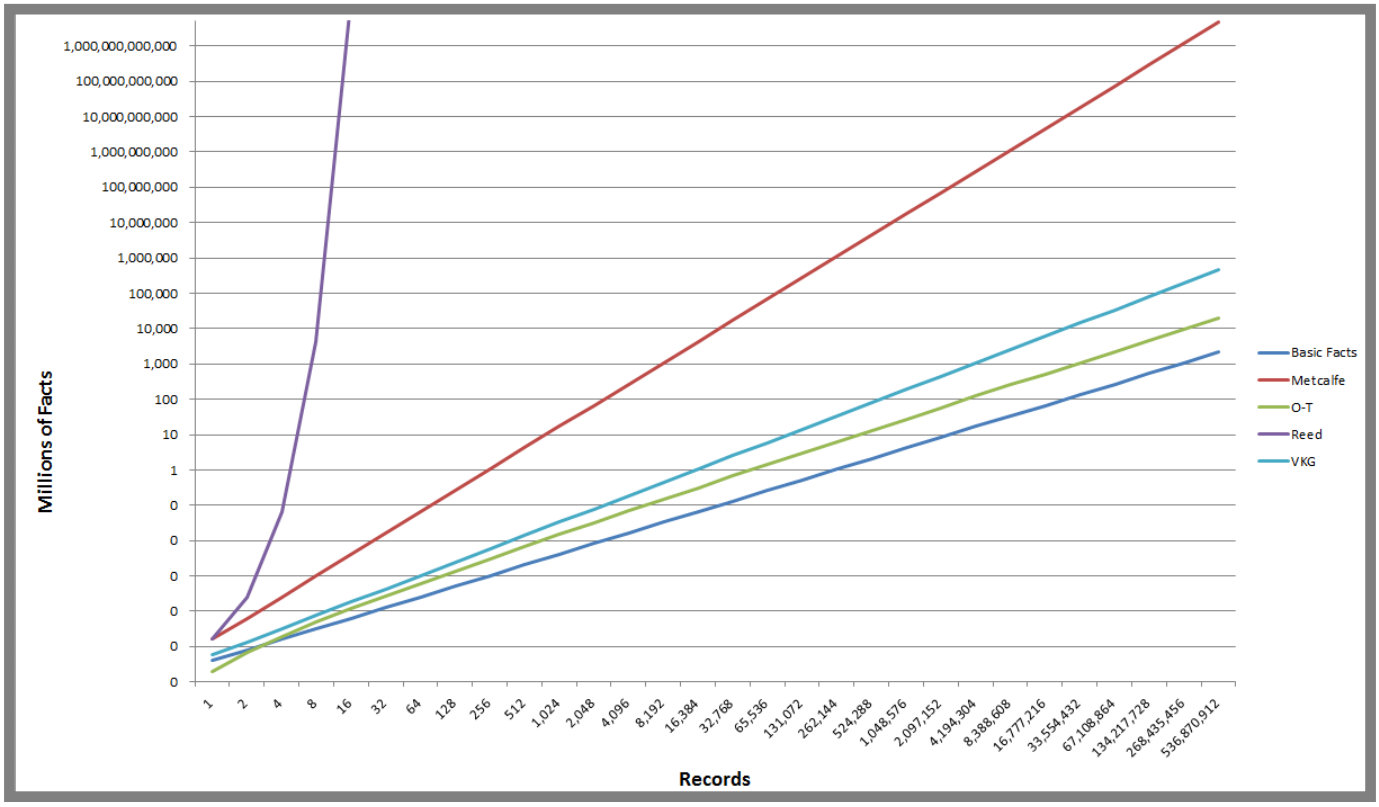

Figure 3. Knowledge Network Estimates

On a logarithmic scale, the O-T (Briscoe-Odlyzko-Tilly) and VKG formulations appear only marginally better than the Basic Facts base case, but that is only due to the swamping effects of the unrealistic growth multipliers. We'll get into this more tomorrow.

Preview of Next Part

The next part forthcoming tomorrow will use this foundation to describe the VKG algorithm, and some of implications of its characteristics, all in the context of knowledge networks or graphs.

[1] "Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets," is a quote from Steve Lohr, 2014, "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights," August 17, 2014, *New York Times*, see http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html. Also, as another example of the common 80% estimate for data preparation costs, see http://radar.oreilly.com/2013/09/data-analysis-just-one-component-of-the-data-science-workflow.html.

[2] These two articles 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. The same concerns I had at that time in the current state of analysis was captured 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] These network effect benefits were reportedly a major driver of Theodore Newton Vail's efforts to consolidate the thousands of initial telephone networks in the United States under the banner of the American Telephone & Telegraph (Ma Bell) company.

[4] See, for example, James Hendler and Jennifer Golbeck, 2008. Metcalfe's Law, Web 2.0, and the Semantic Web, in *Web Semantics: Science, Services and Agents on the World Wide Web* 6(1): 14-20; see http://www.cs.umd.edu/~golbeck/downloads/Web20-SW-JWS-webVersion.pdf.

[5] Babak Hodjat and Adam Cheyer, 2003. Evolution of the Laws that Deal with the Utilization of Information Networks, in Masoud Nikravesh, Lotfi A. Zadeh and Janusz Kacprzyk, eds., *Studies in Fuzziness and Soft Computing*, Vol 164/2005, pp. 427-438, Springer, Berlin. See http://www.adam.cheyer.com/papers/KnowledgeNetworks_Formatted.pdf.

[6] For a well-connected network, every node (*n*) connects to every other node (*n*-1), which gives us *n**(*n*-1) or (n^2 – *n*). Working this out, two nodes have two connections (2*2 – 2), three nodes have six connections (3*3 – 3) and the expression converges on the square of '*n*' for larger values of 'n', *e.g.*, (100*100 – 100) is 99% of (100*100). This convergence at larger number is the basis for the exponential simplification, 2^n. Most of the other 'laws' stated herein are simplifications in a similar manner.

[7] See, for example, Sara F. Peralta, 2011. Moore's Law, Metcalfe's Law, and the Dot Com Bubble, November 27, 2011, see https://storify.com/sarafperalta/moore-s-law-metcalfe-s-law-bubble. Also see [10].

[8] David P. Reed, 1999. That Sneaky Exponential—Beyond Metcalfe's Law to the Power of Community Building, August 27, 1999; online at http://www.reed.com/dpr/locus/gfn/reedslaw.html. For original version, see http://contextmag.com/archives/199903/digitalstrategyreedslaw.asp. Like Metcalfe, at smaller numbers the actual formula is 2^n -*n* - 1, which rapidly converges to 2^n.

[9] However, one group has published an alternative formulation consistent with the Reed approach; see Kalevi Kilkki, and Matti Kalervo, 2004. KK-law for Group Forming Services, in *XVth International Symposium on Services and Local Access*, Edinburgh, March 2004. See http://kotisivukone.fi/files/50ajatelmaa.ajatukset.fi/tiedostot/Others/kilkki_kk-law.pdf.

[10] 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.

[11] See, for example, https://www.princeton.edu/~achaney/tmve/wiki100k/docs/Zipf_s_law.html.

[12] In specific terms, each "fact" in our knowledge base is an assertion, represented as an RDF triple statement. Because some of these assertions may not, in fact, be true, the use of "fact" does not imply universal truthfulness. Rather, an assertion that passes current tests for logic, coherency, consistency or completeness is what is retained, even though its truthfulness is not certain. Therefore, separate adjustment factors (parameters) need to be applied to address accuracy tests.

[13] We adjusted the scale further to reduce the exponential absurdity of the Reed approach by manually shifting the scale downward. As a result, the Reed approach exits the chart rather quickly, heading straight up.

PDF generated by *AI3:::Adaptive Information* blog