Chapter 6 in in Collaborative Computational Technologies for Biomedical Research, edited by Sean Ekins, Maggie A.Z. Hupcey, and Antony J. Williams, Wiley, 2010, 99-112

CONSISTENT PATTERNS IN LARGE SCALE COLLABORATION

Robin W. Spencer

INTRODUCTION

No one will dispute the motivation for this book: without significant change in the rate of new drug approvals over decades [1] despite steeply rising expenditure and knowledge, no stone should be left unturned as we seek higher productivity in health care. Now that internet and storage technologies have reduced the costs of moving and storing information by orders of magnitude, surely we are poised for a positive revolution based on facile collaboration? Yet we are scientists, and a good dose of data is always healthy to calibrate our expectations and help us set strategy. In this chapter we will see that large-scale voluntary collaboration systems show remarkably consistent patterns in contributors' behavior, within scientific collaborations in the pharmaceutical industry and extending to every other company and industry examined. This behavior has all the signatures of a power law, driven by positive feedback from the individual and the group, and with a "long tail" such that approximately half of all contributions come from people who contribute only once to any given campaign. Interestingly the evidence also suggests that networks of acquaintanceship are not an essential driving force which makes us revise our concept of "community". Finally we review the data not just for collaborative idea generation, but for collaborative evaluation and decision-making, and see that the most popular methods are prone to strong bias by minority views.

BACKGROUND

From late 2005 to 2010 I created and then managed the "Idea Farm", an on-line collaborative problem-solving system within Pfizer Inc., the world's largest pharmaceutical firm. The underlying model was the campaign, or challenge, in which a business need is identified with a specific sponsor, the problem or opportunity is reframed for the online medium, then in the "diverge" phase, broadcast (usually via email) to a large diverse audience who then may contribute using an easy-to-use system designed to support the challenge model [2]. In the subsequent "converge" phase, the entries are collected, organized, built upon (by the crowd and/or an assigned review team), evaluated, trimmed, and decisions made on implementation. The challenge model also underpins Innocentive, DARPA challenges (e.g. robot vehicles crossing the desert), X-Prizes, the Netflix Prize, and many more [3]. Arguably the first and most successful challenge was the Longitude Problem, in which the late 18th century Parliament and the Admiralty sponsored an apparently impossible problem to which John Harrison, an unknown clockmaker from the north of England, dedicated his life and which he won by inventing the marine chronometer [4]. If we take innovation to consist of inspiration (the stereotypical "aha"), followed by invention (the proof of concept), followed by implementation (the scaling of the invention and dissemination to its customers), there is no question that implementation is the most lengthy, costly, and likely to fail [5]. The challenge model succeeds because it addresses this at the outset by selecting only those challenges where a serious need is matched by a serious and specific individual who already has the mandate and resources to address the problem, and who will accept input

from a large and diverse audience to get a better and faster solution [6]. While accessible throughout the corporation (secure behind its firewall), the sponsorship of the Idea Farm in Pfizer R&D resulted in a majority of the campaigns involving scientific subjects. The baseline data consist of over 200 campaigns and 3000 separate authors of over 12,000 ideas, supplemented by anonymized datasets from colleagues in similar roles at other large corporations [7].

THE LONG TAIL OF COLLABORATION

"What was the average number of ideas per contributor last year?" is an innocent and reasonable question that has no good answer. It has an answer (typically around 2) , but it is not a good answer because the question implies, incorrectly, that the distribution is somewhat like a bell curve: if the average is 2 ideas per person, there probably were fewer people that put in 0 or 1, and fewer that put in 5 or 10, right? Just like if the average height is 5 foot 7, there should be fewer people 3 feet tall or 10 feet tall? Very wrong. Figure 1 is a rank-frequency plot of over four years' worth of ideas and comments entered into the Idea Farm, where the left-most author (rank #1) put in about 700 entries, and authors ranked 2000-4000 put in one each. A straight line on a log-log plot is a power law; power law distributions are prevalent in natural and human situations [7] but nearly ignored in statistics courses and textbooks. For power law distributions, "average" makes no sense (there is no peak) and the range of values can be enormous. In general, events which are mutually independent (the flipping of a "coin with no memory") will produce Gaussian or normal distributions, while events which are mutually dependent will produce power laws. Avalanches, earthquakes, salaries,

network connectivities all follow power laws, and a strong case can be made that the 2008-2009 financial collapse was due in part to our financial systems' underappreciation of the long tail of this distribution [8]. Figure 1 shows just what a "long tail" of a power law consists of: those 3000 people at the lower right (rank 1000 to 4000) who put in just one, two, or three entries each.

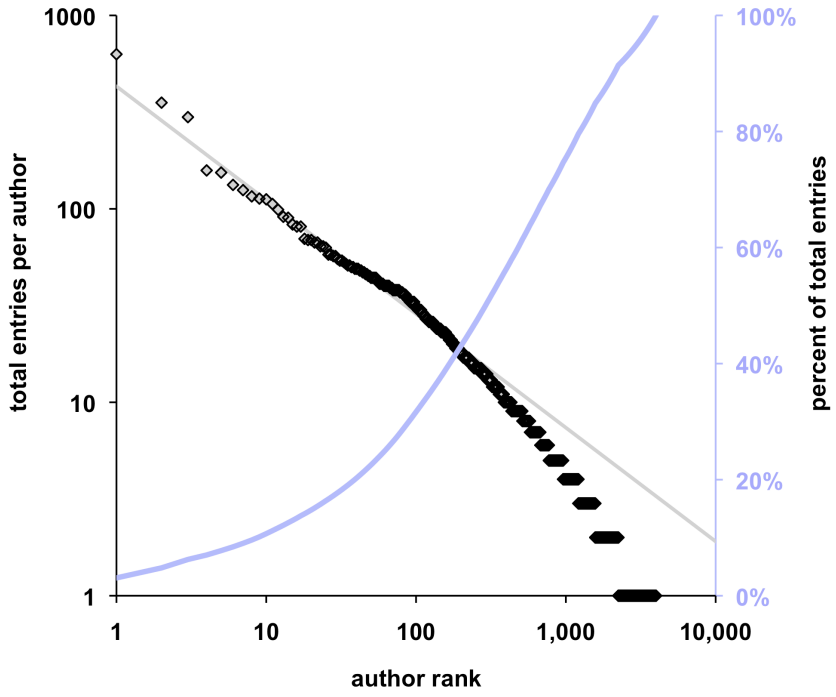


Figure 1: Rank-frequency plot of all ideas and comments submitted to the Pfizer Idea Farm, 2006-2010: 4004 authors, 20,505 entries. Gray line: power law with alpha = 2.7. Overlay curve and right axis: cumulative percent of all entries.

The importance of the tail in a power law phenomenon is the subject of Chris Anderson's eponymous book [9], where he describes how internet technologies, by reducing transaction costs nearly to zero compared to brick-and-mortar stores, enabled Amazon and iTunes to extend the reach of book and music retail to orders of magnitude

more content and consumers than had been previously feasible – to their notable profit and market dominance benefit.

Figure 1 is not unique to Pfizer or even pharmaceuticals or scientific problem solving; Figure 2 shows Pfizer's data (just ideas, not ideas and comments) with data from Cargill, a huge multinational agribusiness corporation. Both companies have secure firewalls with no contact or commonality of people, business needs, challenges, demographics, or cultures, yet their statistics of participation are indistinguishable. This follows for every case we have examined, and at all scales: individual challenges give the same plots with the same slope [10]. This is strong support for an underlying power-law mechanism since it is the only distribution that is scale-free [7].

There is every reason to expect these power law properties to extend to every type of large online collaboration, in part because of the diversity and number of our private-sector datasets [10], and because the contributions to Wikipedia, perhaps the largest open collaborative intellectual effort of all time, follow the same pattern [20].

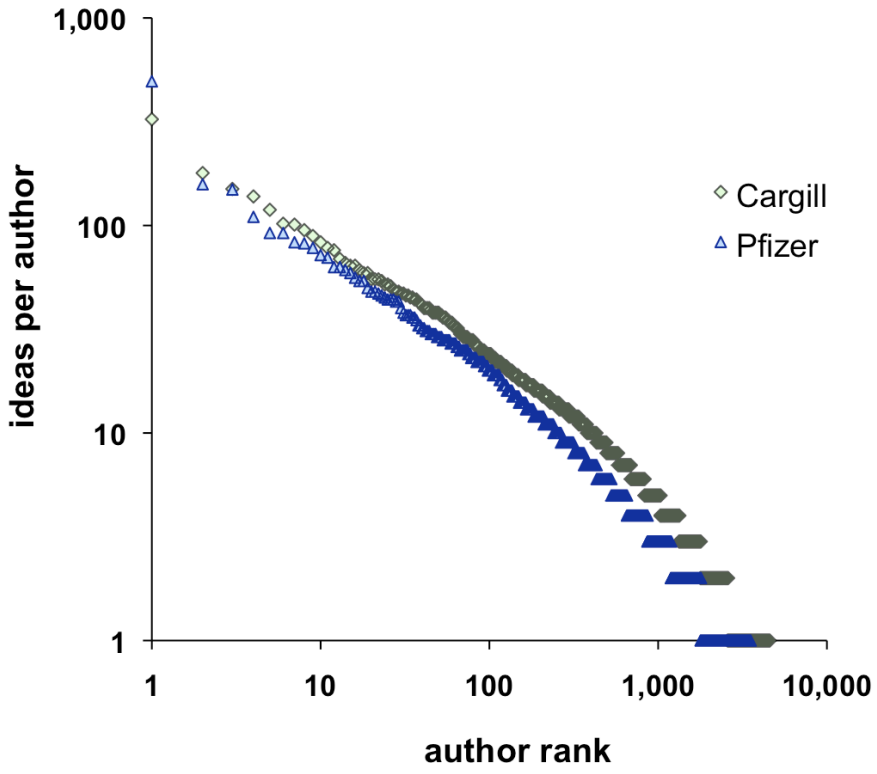


Figure 2 : Rank-frequency plot of ideas submitted to the Pfizer Idea Farm (triangles) and a similar system in a large Cargill business unit (diamonds).

THE VALUE OF AN IDEA

Whether or not the power law property matters depends on the value of what we're measuring, specifically whether the ideas from the "head" (those relatively few people who put in many ideas each) are more or less valuable than those from the "tail" (those many people who put in very few ideas each). Figure 3 suggests three general possibilities, where the ideas are counted in the same order as the participation level of their authors, i.e. ideas from the most prolific authors at the left, and from the occasional authors to the right.

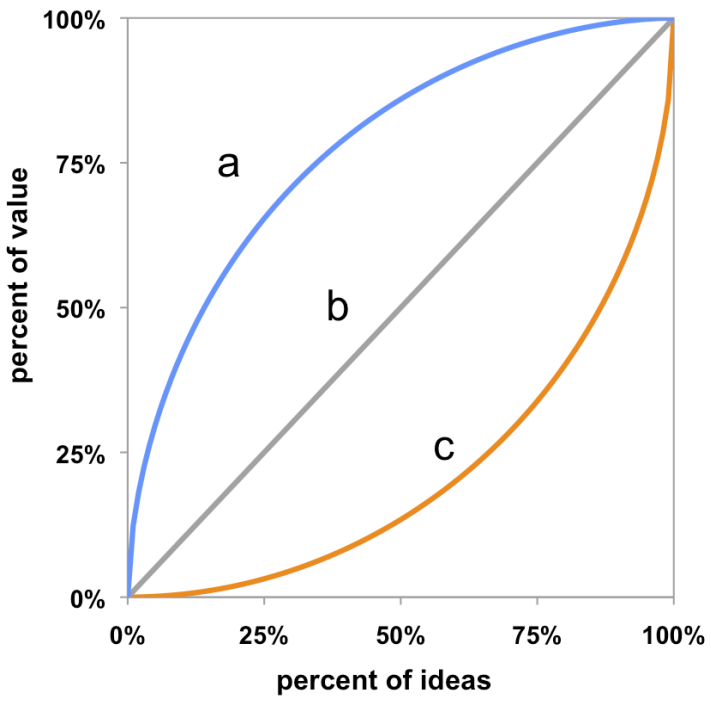


Figure 3 : Models of cumulative value vs cumulative quantity, where quantity (horizontal axis) is ordered by author rank (as in Figures 1,2). (a) "Head" participants have better ideas, (b) all ideas are [probabilistically] equal, and (c) "tail" participants have better ideas.

In case (a), the "idea people" are not only prolific, their ideas are better. If there is some sort of expertise or talent for idea generation, or if people with more talent also have more confidence resulting in higher participation, this is what we might expect. It would be a case of an "80-20" rule where a minority dominates in quantity and quality. On the other hand, a case can be made for (c), where the ideas from the rare participants should be better: there is good evidence that teams get tired and less effective over time and need external stimuli [11], and Gary Hamel makes a strong case that value comes by "listening to the periphery", those voices seldom heard by dint of corporate culture, geography, or generational deafness [12].

Our data, while not as complete as for the power law itself, are consistent and provocative. Figure 4 shows the results from four large Pfizer challenges in which semiquantitative estimates of idea value were available. Importantly in all cases entry value was assigned by the review team established by the campaign sponsor, and judged by criteria agreed in advance (typically along dimensions of technical feasibility, potential market value or cost or time reduction, competitive advantage, and IP risk) – in other words, value was estimated by those who would benefit by success and be involved in implementation. Ideas rated low by such a team have essentially no chance to be implemented, and so however intrinsically brilliant, have no value: a harsh pragmatic reality. Such teams usually begin with a high-medium-low binning before any ideas are excluded; Figure 4 shows the results with high=10, medium=4, and low=0 points, though the weighting factors made no qualitative difference to the results.

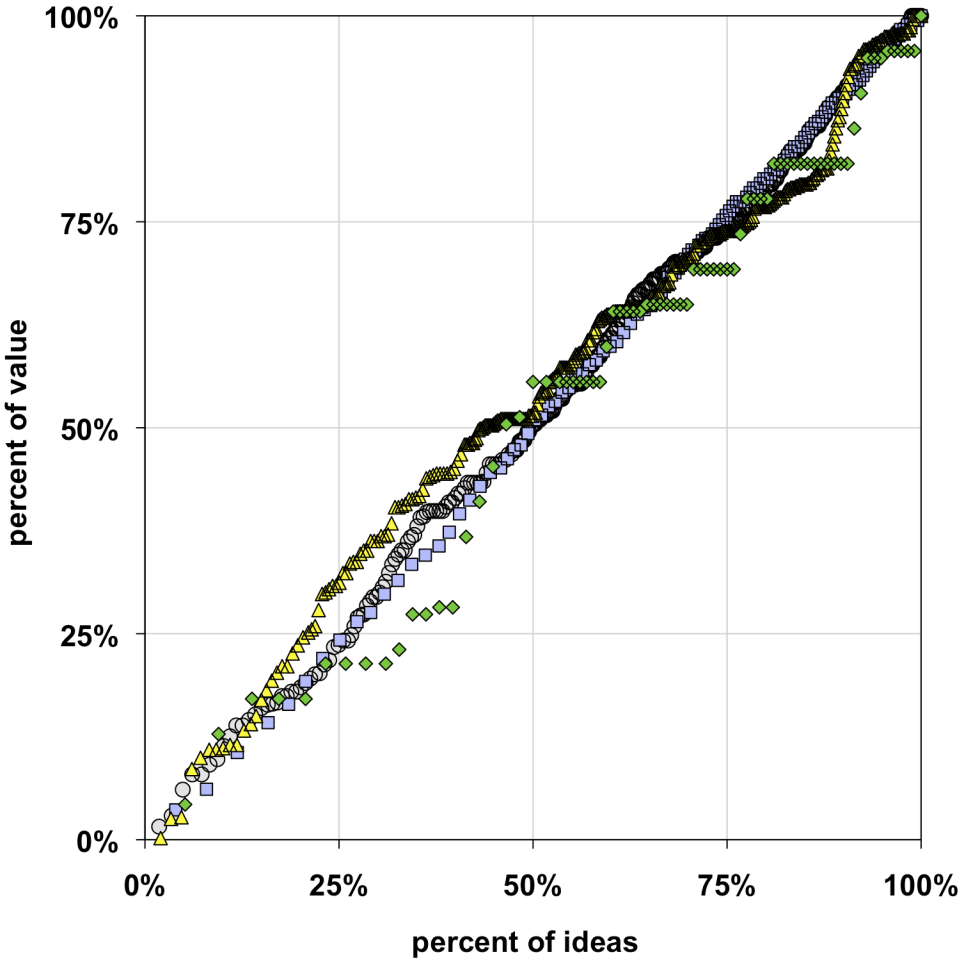


Figure 4: Cumulative value vs cumulative quantity from four large Pfizer campaigns: (▲) an all-R&D campaign seeking non-traditional, marketable IP, (●) a challenge to reduce operating costs for a mobile sales force, (■) a process-improvement challenge to reduce time and complexity in clinical document preparation, and (◆) a scientific-medical challenge for additional indications for an existing drug.

This is a very interesting result, supporting neither above hypothesis but rather suggesting that idea value is independent of whether the author is prolific or occasional. In other words, if 1 in 100 ideas is valuable (for example), then we might expect one valuable idea from the 3 people who put in 50, 30, and 20 each, and also might expect one valuable idea from the 100 people who only put in one each. Note how accurately this parallels

the value proposition in iTunes or Amazon's Kindle bookstore, where songs cost about 99 cents and books cost $10, roughly constant from best-sellers to the most obscure titles.

Now we can return to the overlap of Figure 1, the cumulative area under the log-log graph. Not only does this represent the cumulative number of entries, it represents the cumulative value. About half the value is contributed by authors 1-300, and the other half by authors 301-4000. If your reaction is to think "I want those top 300!", you are missing the opportunity of large-scale collaboration in three important ways. First, exactly which 300 you need is going to change for every given business problem. Innovation must be specific and purposeful; calls for "we only want big game-changing ideas" are guaranteed to fail [5], and so successful campaigns are quite content-specific and useful contributions draw on deep personal expertise and experience. Secondly, traditional teams become dysfunctional beyond a dozen or so participants [13]. Even scheduling meetings for a team of 20 becomes infeasible. If you fall back to the idea of "top 10" for a team, Figure 1 tells us that you will knowingly miss 90% of the value you could have had. If your organization doesn't have 4000 people in it, that's still true, they just don't all work for you. Thirdly, and optimistically, recall the lessons of Chris Anderson: don't run from the long tail, exploit it. Systems designed to facilitate, run campaigns, and manage evaluation and next-steps are readily available [2], and the cost is not the system, it's the opportunity lost by ignoring the value in the tail (which you now know how to predict).

In fact the tail value is greater for individual campaigns than Figure 1 suggests. The only exceptions to power law behavior seen to date are from non-voluntary campaigns. The electronic tools for mass collaboration work equally well in a

"command" situation; for example, it is very productive to have a half-day meeting with several background presentations followed by a "flash event" in which every member of the audience is exhorted to spend the next 15 minutes writing down 4 ideas for how their work team could support the presented proposal [13]. In these cases the result is not a power law distribution [10] but closer to a Gaussian; on a rank-frequency plot the tail drops quickly. The Pfizer data is an aggregate of many campaigns including large involuntary ones of this type, which probably accounts for the deviation from power law at the bottom right. When large voluntary individual campaigns alone are considered, the tail extends farther [10, figure 5a], with the consequence that fully half of the total entries (and therefore value) come from people who only ever contribute once [10, figure 8].

COMMUNITIES ?

We have seen that voluntary, large-scale, collaborative challenges on scientific topics are feasible, sustainable, technically well understood, and that a great deal of the value derived comes from the occasional contributors. But are these really "communities," or is that word becoming overworked, in the same way that calling a stranger who accesses your blog or photos a "friend" doesn't make them one. It's more than a semantic quibble if our beliefs affect our strategies for attracting new participants or rewarding and recognizing past contributors.

We have one relevant dataset, but it's objective and large scale. For years Pfizer, like many companies, has had a link on its public website, saying in effect, "send us your ideas to improve our offerings." Figure 5 shows the familiar rank-frequency plot for several years' of this activity; again, it is an excellent power law. What this dataset has in

common with the others is that it's from a large scale voluntary process, seeking new ideas and concepts for business purposes. Where it differs is that for intellectual property and legal reasons, the process has been implemented as a "drop box", in which contributor's identities are not accessible to each other and there is no possible commenting or cross-contributor collaboration (a hallmark of the internal challenges). These contributors cannot by any stretch be called a community, because they cannot know or communicate with each other. And yet we have a power law signature, including the same exponent (alpha = 2.7 ± 0.3, [10]) as all others observed.

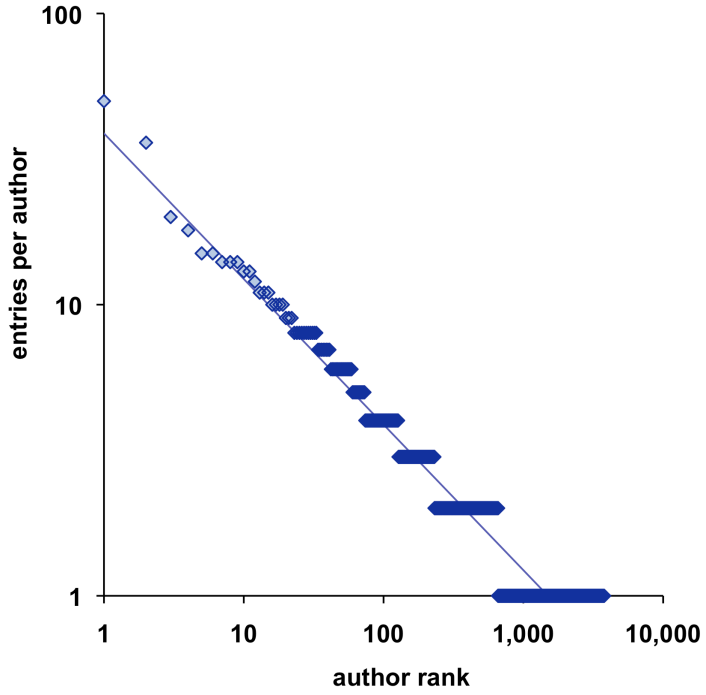


Figure 5: Rank-vs-frequency plot for unsolicited open website suggestions. The line has exponent alpha = 3.

Figure 5 refutes the hypothesis that our power laws might derive from a network effect. It is well established that human networks show power law statistics in their

connectivity [15], and it would be reasonable to suppose that our observations somehow derive their statistics from a driving force dependent on a network: for example, I'm more likely to contribute if people in my social network contribute. It remains perfectly possible that such network effects could amplify the contributions to a challenge, but Figure 5 shows that something more intrinsic, more local is going on. A source of positive feedback is the most likely origin of the power laws [7], and simulations suggest that feedback comes approximately half from one's own behavior ("I put in an idea last week, it wasn't hard or scary, I'll probably do it again.") and half from general observations of others ("Other people are doing this, I'll give it a try.") [10]. The difference between general ("other people") and specific ("people I know and trust") is arguably the difference between a collaboration process and a community-driven process.

With "community" now shown to be a tenuous concept, we have to consider how to advertise our campaigns and induce people to contribute. We can certainly hope that people will tell all their friends, but cannot rely on it, and have data suggesting instead that contribution is more likely a private choice. As manager and facilitator for hundreds of challenges this is not surprising: almost without exception, announcements of a new broad challenge that depend on propagated emails will fail: first because the emails simply don't get sent, or second, because they're sent with a generic title ("Please Read", or "On Behalf Of") which does nothing to convey the content or opportunity. In a world of spam and information overload, this is a bucket of cold water.

MOTIVATION AND SUSTAINABILITY

If we can't expect true community behavior, and if the specificity of our business needs is such that each campaign will uniquely interest different people, how can we make large scale collaboration work? My role at Pfizer brought me into a true community of peers from other companies, brought together face-to-face in vendor-sponsored user groups. There are definitely best and worst-practices, learned over and over, for example:

- Do not offer tangible prizes or rewards. Especially for cutting-edge scientific challenges, the participants you need are probably well paid and not particularly enthused by another tee shirt, coffee cup, or $100 voucher. There is intriguing literature that, in fact, monetizing an otherwise altruistic bargain will decrease participation [16]. If that's not enough, offering tangible rewards comes at significant cost: Who will get the prize? (Let's call a meeting…). Who has the prize budget? Just don't do it unless you're not prepared to make a full business of it (for example, like Innocentive's prizes which may typically be in the $5,000 to $40,000 range).
- Do offer recognition but watch for overload. Absolutely recognize contributors when campaign results are known; every organization has appropriate newsletters for this. But beware of the cynicism that follows from too many employee-of-the-month type programs [17]. It's not kindergarten, not everyone gets a star.
- Highlight based on quality, not quantity. Since most of your campaign value will come from people who only ever put in 1, 2, or 3 contributions, don't cut off the tail by hyping the high contributors and implicity offending the rare ones. Don't set up a "reputation" system based on mouse clicks rather than serious content. Do highlight contributors but based on quality, not quantity.

- Remember Herzberg. A generation ago, Herzberg studied employees' motivators and demotivators; his article [18] has been the most-requested reprint from the Harvard Business Review. Even more important than recognition is to make the task serious and real; people seek achievement and responsibility. In other words, the challenges you pose must *matter*, and it must be clear how they matter and to whom. Never pose toy challenges or ones that address minor issues; it devalues the entire program. Equally important is to assure that your collaboration system avoids the Herzberg demotivators (or "hygiene factors"), principal of which is the perception of unfair or inappropriate policies and bureaucracy. In other words, if you want voluntary help, don't make the contributor suffer through three pages of legal caveats or a picky survey, and make the challenge about something known to be important to the sponsor and, perhaps altruistically, to the contributor.

COLLABORATIVE EVALUATION

Soliciting and collecting ideas is only the divergent half of a campaign; the convergence process of evaluation and decision must follow if there is to be implementation. For typical departmental-scale campaigns in which entered ideas number in the dozens to hundreds, a review team appointed by the original project sponsor is very effective, because it taps directly into the organization's norms for project responsibility and funding. However, when entries approach the thousands it may be useful to enlist the "crowd" to assist in their evaluation.

But the data suggest caution. Figure 6 illustrates how we need to be aware of the possibility that crowd evaluations, however democratic and open in intent, may be driven

by small minorities. The types of data in Figure 6 appear to be in order of difficulty or knowledge-required (i.e. a 5-star vote takes less effort or know-how than typing in an original contribution), which suggests an important possibility: that the easier or less content-rich the task, the more it is likely to be driven by an active minority of participants. This is perhaps counterintuitive: "make it easy" would seem to encourage a more democratic, representative outcome. But the datasets behind Figure 6 are so large that we must take very seriously the possibility that "make it hard and specific" is the better way to assure a broader source of input.

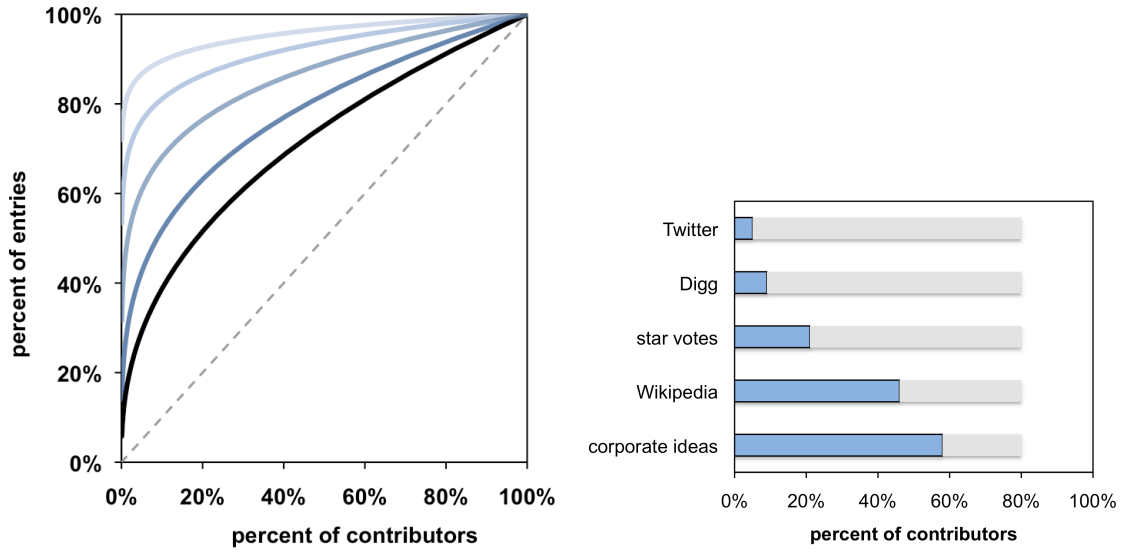


Figure 6: The exponent of power laws reflects the distribution of participation. (a, left): Curves are Newman's equation 28 [7], for (top to bottom) alpha = 2.05, 2.1, 2.2, 2.4, 2.7, which are approximately the exponents for Twitter entries [19], Digg promote-demote voting [20], five-star voting (this work and [10]), Wikipedia edits [20], and corporate ideas (this work and [10]). Dashed gray line, if all participants contributed equally. (b, right): Slice through Figure 6a, illustrating how many people contribute 80% of the content.

Of equal concern are the very large scale observations of 5-star voting at Amazon [21], namely that it is biased, compressed (with an average of 4.4 out of 5 stars), and prone to follower behavior that drives to extreme opinions rather than balance. Consistent with

the observation of Figure 6, the authors recommend making the online book review process more difficult, rather than less, to achieve better quality and balance.

CONCLUSIONS

Multiple large datasets from diverse private and public sources show that contributions to large scale voluntary collaboration campaigns (including scientific challenges) generally follow a power law, and with an exponent considerably higher (alpha = 2.7 ± 0.3) than "easier" tasks (Twitter, Digg, 5-star rating; alpha ca. 2 to 2.2). The consequence is that these campaigns depend for a majority of their content on a "long tail" of people who contribute only a couple of ideas each. Because power laws are scale free this generalization applies to small as well as global-scale campaigns. The phenomenon may benefit from, but does not depend on, social networks because blinded "drop box" challenges have the same signature. Thus rather than speak of "communities" we would be more accurate to refer to "personal responses to a particular challenge." To encourage participation we should respect our contributors as individuals, recognize quality over quantity, and remember the strong motivation of contributing to real work that makes a difference. Large-scale collaborative evaluation of options is more problematic, since the data for popular techniques like promote-demote and 5-star voting reveal a potential for considerable bias and dominance of minority opinions.

Acknowledgements

The author, recently retired as Senior Research Fellow at Pfizer, Inc., wishes to thank Steve Street for his support and courage to embrace change, as well as Tim Woods at

Imaginatik, Doug Phillips at Pfizer, Anne Rogers and Kurt Detloff at Cargill for sharing anonymized data. He may be contacted at spencerrw@alum.mit.edu .

REFERENCES

[1] Summary of NDA Approvals & Receipts, 1938 to the present, at http://www.fda.gov/AboutFDA/WhatWeDo/History/ProductRegulation/SummaryofNDAApprovalsReceipts1938tothepresent/default.htm, accessed 5 July 2010.

[2] The system behind the Pfizer Idea Farm is Idea Central® from Imaginatik PLC.

[3] See http://www.darpa.mil/grandchallenge/index.asp, http://www.xprize.org/, http://www.netflixprize.com/ for examples.

[4] Sobel, D. Longitude. London: Fourth Estate, 1995.

[5] Drucker, P. Innovation and Entrepreneurship. New York: HarperBusiness, 1985, ch. I.2.

[6] Suroweicki, J. The Wisdom of Crowds, New York: Doubleday, 2004.

[7] Newman, MEJ. Power laws, Pareto distributions and Zipf's law, at http://arxiv.org/abs/cond-mat/0412004v3, 2004.

[8] Olmerod, P. Why Most Things Fail. New York: Pantheon, 2005, p.173-179.

[9] Anderson, C. The Long Tail, New York: Hyperion, 2006.

[10] Spencer R, Woods T. The Long Tail of Idea Generation. International Journal of Innovation Science, in press.

[11] Katz R, Allen TJ, Investigating the NIH Syndrome. In: Tushman M, and Moore, M, editors. Readings in the Management of Innovation, New York: HarperBusiness, 2nd ed, 1988, p. 293-309.

[12] Hamel, G. Leading the Revolution, Cambridge: Harvard Business School Press, 2000, p. 261-264.

[13] Brooks, F. The Mythical Man-Month, New York: Addison-Wesley, 1995.

[14] Spencer, R. Innovation by the Side Door. Research-Technology Management, 2007; 50:5:10-12.

[15] Barabasi A-L, Albert R. Emergence of Scaling in Random Networks. Science 1999; 286:509-512.

[16] Bowles, S. When Economic Incentives Backfire, accessed at http://hbr.org/2009/03/when-economic-incentives-backfire/ar/1

[17] Deming, WE. Out of the Crisis. Cambridge: MIT Press, 2000.

[18] Herzberg, F. One more time: How do you motivate employees? Harvard Business Review 1987; 65:5:109-120.

[19] Piskorski, M. Networks as covers: Evidence from an on-line social network. Working Paper, Harvard Business School, at http://www.iq.harvard.edu/blog/netgov/2009/06/hbs_research_twitter_oligarchy.html

[20] Wilkinson, D. Strong regularities in online peer production. In: Proceedings of the 2008 ACM Conference on E-Commerce, Chicago, IL, July 2008, at http://www.hpl.hp.com/research/scl/papers/regularities/

[21] Wu F, Huberman BA. How public opinion forms, Social Computing Lab, HP Labs, Palo Alto, at http://www.hpl.hp.com/research/scl/papers/howopinions/wine.pdf, and Kostakos, V. Is the crowd's wisdom biased? A quantitative assessment of three online communities , at http://arxiv.org/pdf/0909.0237