Frank Schweitzer:
Limits to Scaling – and How to Move beyond Them
in: F. Dombois, J. Harboe (eds.): Too Big To Scale – On Scaling Space, Numbers, Time and Energy,
Scheidegger&Spiess (2017), pp. 161-176

Limits to Scaling – and How to Move beyond Them

Frank Schweitzer
Chair of Systems Design, ETH Zurich
Weinbergstrasse 58, 8092 Zurich, Switzerland
fschweitzer@ethz.ch

Scaling - a systemic property

Small and large systems In *engineering*, the term "scaling" may be used to describe a relation between *objects* and their *models.* In the *exact sciences*, however, scaling is a systemic property, i.e. it relates to the system itself. Scaling describes how certain system properties change conditional on others, most notably the system size.

Taking the example of a social network, the quantity of interest could be the number of friends someone has in this social network. Then, the corresponding question would be how the average number of friends scales with the size of the system, i.e., the total number of members, N, of the social network. Obviously, if N is ten, the the maximum number of friends is bound to ten, and the average number should be even less. But what happens if N equals one thousand? Does the increase in the *potential* number of friends also leads to an increase in the *actual* number of friends? And how does this scale if N approaches ten million?

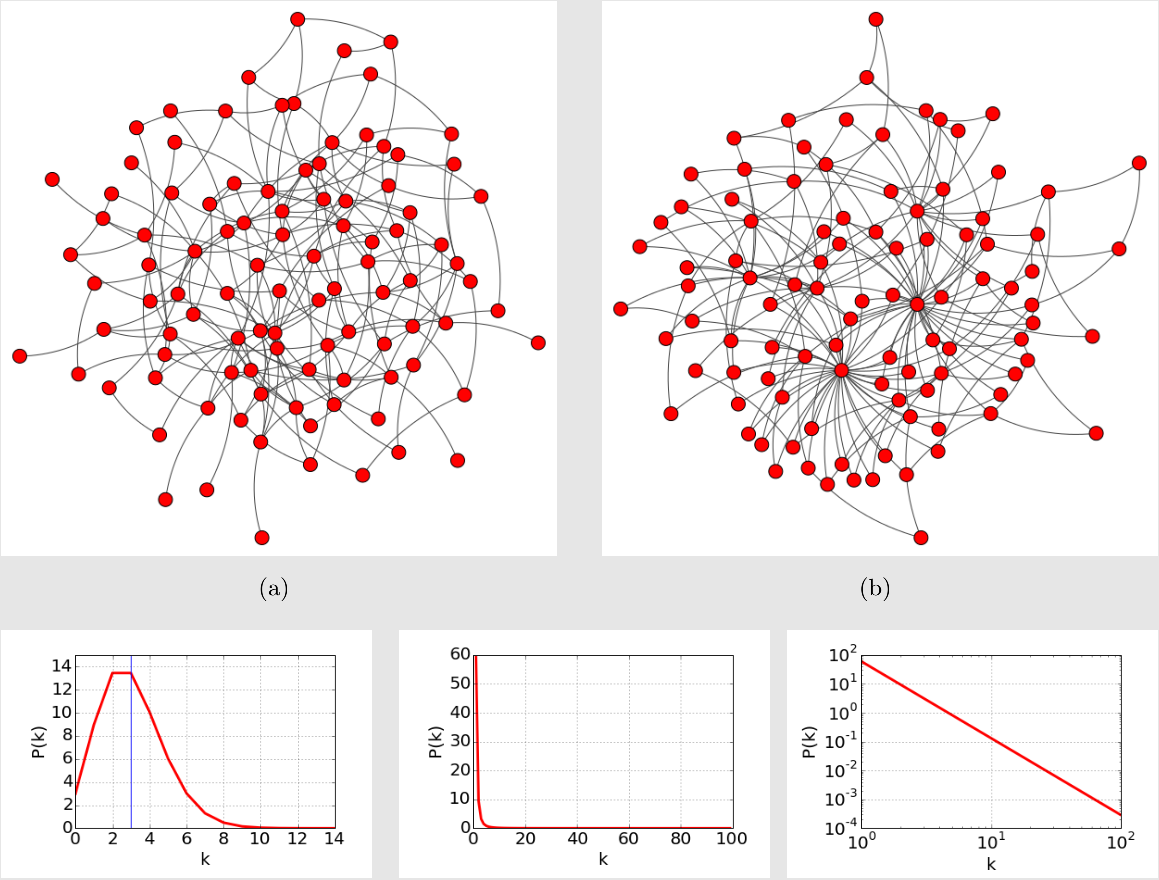
Discrete and continous measures The example at hand uses a *discrete* measure of system size, N. But scaling can be also applied if *continuous* measures for the system size are used, for example the volume V or the mass M of a physical system, These should be *extensive* measures, i.e. they increase with system size, as opposed to *intensive* measures which are independent of size. In fact, these extensive measures are in the limit of large numbers, e.g. for $N \to 10^{23}$ which is the number of atoms in one mole of a given substance, all related by simple linear transformations: $N \propto V$, where the proportionality constant is given by the particle density (which is the intensive measure), or $N \propto M$, with the molar mass of that substance as the proportionality constant.

Simple transformations Such linear transformations are quite common in *geometric* scaling, e.g. a system of length X, width Y and height Z can be scaled to a system with $X' \to aX$, $Y' \to bY$, $Z' \to cZ$. *Isotropic* scaling means that the proportionality constants a, b, c are all the same, otherwise *anisotropic* scaling is observed. If these constants are larger than 1, the system is *enlarged* (or stretched), otherwise the system is *contracted.*

Interesting transformations can be observed if some constants are larger and others are less than, or equal to, 1, or if instead of constants *functions* are used. An example of a *linear function*, also known as *shear mapping*, would be $X' \to aX + bY + cZ$. The famous biologist D'Arcy Wentworth Thompson has used such transformations in his book *On Growth and Form* (1917) to describe the body shape of certain related biological species. An example of a *non-linear function* would be $X' \to (aX)^2(bY)$.

The discussion of such transformation, or *deformations*, is mostly restricted to the geometric shape of a system. Instead, we are interested in a very different question, namely how system properties change with the *size of the system*. To refer again to the physical system: is it possible that a very different system behavior, e.g. a phase transition from vapor to water can be induced/prevented if we just change the *size* of the system and keep everything else constant?

No scaling - but a natural scale Certain systems are characterized by a natural scale that remains the same even if the system size increases. For example, the average number of offspring per human remains roughly the same irrespective of the size of the system (i.e. the number of mating options). Such systems are often described by the so-called normal distribution, $P(x) \propto \mathcal{N}(\mu, \sigma^2)$, where the location parameter μ is given by the mean value (e.g. the average number of offspring) and σ^2 characterizes the variance. The latter is in statistics often denoted as the scale parameter as it describes how spread out the distribution is. For normal distributions, we expect only relatively small deviations from the mean. In plain words: systems with natural scales are very predictable with respect to that scale, or quantity.



(a)

(b)

(c)

(d)

(e)

A fundamental difference Networks (a) and (b) each consist of $N = 100$ nodes representing users and links depicting their interactions. While the networks at first sight seem to be similar at first glance, they very different with respect to the number of links, denoted as *degree* k, of each user. The degree distribution $P(k)$ gives us the frequency of finding a given degree k in the network. Network (a) is a so-called *random* network characterized by a Poisson distribution $P(k)$, shown in (c), which for large networks converges to the normal distribution $\mathcal{N}(\mu, \sigma^2)$. I.e., the degree distribution has a well defined mean, 3 in this case, and a rather small variance. This would not change if we considered $N = 10.000$, i.e. the mean degree would still be 3, while outliers may have 10 or so. Network (b), on the other hand, is a so-called *scale-free* network characterized by a power-law distribution $P(k) \propto k^{-\gamma}$. It is shown in (d), using the same linear scale as in (c) to allow comparison, and in (e) using a log-log scale. The straight line in (e) indicates that there is a non-negligible probability of finding few users with a very large degree. Hence, if we increase the network to $N = 10.000$, we see a larger number of users with degrees much larger than 3, maybe with degree 300 or 3000, while in (c) this probability can be safely neglected.

No natural scale - but scale-free Other systems are scale-free in the sense that no characteristic scale exists. An example from social communication: what is the typical time interval x at which we reply to a mobile phone call, a post in a chat, a written letter? There is no such typical time lapse. Instead, we find that any possible value x, from minutes to months, occurs with a well-defined frequency that is described by a so-called scale-free distribution, also referred to as a power law, $P(x) \propto x^{-\gamma}$. Scale-free refers to the fact that if we replaced x by cx, where c is a scaling constant, i.e. if we measured x in days instead of seconds, we still have $P(cx) \propto x^{\gamma}$. Just the proportionality factor is different and now involves the constant $c^{-\gamma}$. Hence, the distribution $P(x)$ is scale invariant. It depends on only one parameter, the scaling exponent γ, to determine how often a certain value x appears

Remarkably, for human communication this exponent is $\gamma = 1.5$, irrespective of whether we analyze communication in chat rooms, over mobile phones, or the correspondence of Darwin or of Einstein. Hence, this scale-free distribution captures the very essence of our human communicative behavior. Depending on the value of γ, sometimes not even a mean value μ or a variance σ^2 is defined. In plain words: in scale-free systems, even rare events can happen with a non-negligible probability.

Not scale-free, but very broad In most real cases, the distribution that determines how often a certain value appears is not exactly scale-free, but very broad. An example from scientometrics: How many citations obtains a scientific publication within ten years? The number of citations has no typical value. Most publications never get cited, while a rather small number of publications gets cited several thousand times.

The frequency of a given number of citations follows a *log-normal distribution*, which looks very skewed, like a power-law, for large numbers of citations, but like a normal distribution for very small numbers of citations. In particular, it still has a defined mean value μ and variance σ^2. The citations of publications in *different scientific disciplines* all follow a log-normal distribution, however, they do not follow *the same* log-normal distribution. Instead, their characteristic values μ and σ^2 depend on the discipline and indirectly also on the *size*, i.e. the number of publications in that discipline.

Scaling the distribution Interestingly, we can rescale these citation distributions. I.e., we divide the number of citations x by the average number of citations x_0 received by all publications in that discipline published in the same year. Plotting then the scaled distributions $P(x/x_0)$, we see that the log-normal distributions $P(x)$ from different disciplines all collapse into *one* master curve.

Hence, we have obtained two insights: (i) we verify that the probability to find a paper with a given number of citations is described by the scaled distribution $P(x/x_0)$, irrespective of scientific fields or years, (ii) we have obtained a simple *scaling function* $x \rightarrow x/x_0$ that accounts for the influence of scientific disciplines via the characteristic value x_0.

The scaling function comprises two different sizes: the size of the community representing the scientific discipline and the number of publications in the respective discipline. Remarkably, these "sizes" enter the description not directly, but indirectly. I.e. the scaling function contains this information in a nonlinear, and very compact, manner. Finding such simple yet robust and quite universal scaling functions is one of the ultimate aims of scientific research in the exact sciences.

Scaling in input-output relations Social and economic systems not only have their specific structure and dynamics, they also serve a purpose. That means that we can relate their performance, or output in general, to other variables, notably the system size or other input variables. An example from economics: If we increase the *input* of capital and labor by a factor of two, how would that impact the *output* of the system, e.g. the production of some goods?

The most common case is known as *decreasing returns to scale* (DRS): If we double the input (e.g. labor), we do not get twice the output but *less* because, for example, the administrative costs for hiring and managing a larger workforce have considerably increased. DRS set limits to further growth because beyond a given production size any further increase of production is rendered *unprofitable*. The desired scenario would be *increasing returns to scale* (IRS): If we double the input, we get *more* than twice the output. I.e., the more, the better. Economists argue a great deal about how to turn DRS into IRS.

Emergence of new systemic properties

More is different The above considerations imply that the system properties do not change with scale other than the way defined by the scaling functions. But complex systems, i.e. systems consisting of a large number of strongly interacting elements (usually denoted as agents), also have the property of *emergence*. This describes the sudden occurence of new system qualities once certain critical parameters, known as thresholds or *tipping points*, are reached. Other common terms for this sudden occurrence are *phase transitions* or *sudden regime shifts*.

An example from physics is conductivity, which is a systemic property. A single atom has no conductivity. So, how many atoms do we need to observe something like conductivity? For mercury, one finds that conductivity emerges when the number of atoms increases from a few to up to 100. Below 10, the system has no conductivity, but above 100, its measurable value cannot be distinguished from a macroscopic bulk phase with millions of billions of atoms. In plain words: above a certain system size, the system behaves completely differently.

Critical mass? Sometimes, it is very difficult to know when this critical size is reached. Actually, it is not always the *size* that defines the threshold for the emergence of new system qualities. Quite often, it can also be the interaction strength between system elements. Weak interaction may not lead to emerging phenomena, while strong interaction does.

Many variables involved in determining the tipping point are not fixed from the outside as e.g. boundary conditions, but evolve over time with the dynamics of the system. In a growing system, the system size increases over time until a critical value is reached. However, information, too, can be accumulated over time until it reaches a level that changes the systemic behavior.

That means that we need to know not only the *critical value*, but also the *time* needed to reach this value. This makes it so difficult to predict the *onset* of emergence.

Critical slowing down/ speeding up Once a critical threshold has been reached and a phase transition or a sudden regime shift has occurred at a particular time t_c, we cannot assume that established regularities still hold afterwards. This also implies a breakdown of scaling relations which need to be replaced by other relations describing the new phase, or regime.

Remarkably, however, the dynamics of *approaching* the tipping point of a phase transition follow their own *time-dependent* scaling laws. This is known as *critical slowing down*, but critical speeding up can also occur. It means that, in the vicinity of a tipping point, the system does not behave just randomly, but shows signatures in the dynamics that indicate a coming phase transition. They often contain terms of the form $(t_c - t)^\alpha$ where the scaling exponent α can be positive or negative. For example, in ecology certain processes such as recovery occur at a much lower rate than they do otherwise, whereas in economics fluctuations of prices may considerably increase when approaching tipping points.

Beyond scaling

Change the system Distributions define the probability of a given value, while scaling functions describe how system properties change with system size or with input variables. This allows us to predict the systemic behavior, on the other hand also sets limits to what should be expected from a system – and what not.

However, the most interesting question, for both social and economic systems, is how to get beyond scaling. Because scaling is a systemic property, this implies *changing* the system. Change, on the other hand, is a double-edged sword: social and economic systems are *adaptive systems*. That means that whatever we propose to “improve” the system will result in a response of the system in *both* intended and unintended ways. “Unintended” is not the same as “unforeseeable” – a systemic perspective could indeed help us to better understand the occurrence of unintended consequences. But in most cases, social and economic actors, both on the individual and the institutional level, have incentives enough *not* to adopt such a perspective.

From decreasing to increasing returns to scale Physical production is constrained by material resources, so it naturally obeys decreasing returns to scale. This sets limits to growth and, hence, to profit. To vastly increase the latter, production has to be transformed to obey increasing returns to scale.

This happens in the internet economy and is related to the *network effect:* If the size of a social network is measured by N, the number of users, we have potentially N^2 connections between all users. If we *double* N, we get *four* times more connections, if all of these connections can be utilized. Knowledge production and other forms of innovations crucially depend on such network effects.

Increasing returns to scale are also related to the production of non-rivalrous goods, i.e. goods that can be possessed or used by more than one actor without their value diminishing. Examples are electronic books, music or video files and other type of information related products, that can be easily shared (and sold) without further increasing the production costs.

Hence, the digital economy can obey increasing returns to scale. Also the modern financial industry is able to realize this. While classical stock exchange relies on a limited number of available assets that have to be possessed to be traded, derivatives are financial products that no longer require to possess the so-called "underlying" and thus allow the exponential growth of the financial market.

From decreasing to increasing attention The number of citations is bound by the size of the scientific community that becomes aware of the respective publication, e.g. by browsing scientific journals. While the resource, i.e. the size of the community, cannot be easily increased, the number of publications increases exponentially, which implies potentially less attention per publication.

To turn decreasing into increasing attention, one has to change the system such that new mechanisms generate more awareness for one's own publication. So, scientists launch new ways of self-marketing, preferably in social media, to popularize their work. It is then no longer the scientific quality that drives their increase in citations, it is the impact of social and mass media.

Uninteded consequences The above mentioned changes introduce new feedback cycles in the system, and the system responds to these via other feedback cycles in a nonlinear and scarcely predictable manner. Thus, in addition to the intended changes, there are also – always – unintended consequences if one attempts to overcome the limits set by scaling.

Increasing returns to scale in the Internet economy attract financial investors seeking extra profit. This leads to venture investments and drives financial bubbles, as recently seen with the *dot-com* bubble. New derivatives, originally intended to diversify risk, result in even more risky behavior and eventually in financial crises of global scale. The public advertisement of the latest scientific results in social media, in line with over-simplification and over-selling, leads to a decrease of reputation not just of individual researchers, but of whole scientific disciplines.

A way out? The only way of mitigating unintended consequences is to better understand the system by means of a systemic perspective. There are several limits to scaling which need to be *detected* and *respected.*

Bigger is not better: decreasing returns to scale render further growth unprofitable. And increasing returns to scale directly lead to bursting bubbles. Thus, it makes no sense to just stretch the limits to growth, or to scaling. Moreover, *small is not beautiful*: the occurrence of emergent phenomena beyond certain thresholds may lead to completely different systems and render our scaled-up knowledge from smaller systems useless. But *more is different.* And we need to understand, in which way.