Online Appendix

Online Appendix

Expected Seed Pool Size

I compute the expected seed pool size for a guiding conception, denote its elements 1 and 2, in a random network in which there are N top elements, m bottom elements for each top element, and the network is generated according to the two-step process outlined in the text with p the probability of a link between a top and bottom element in step 2. I will take p to be on the order of $\frac{1}{N}$ or equal to $\frac{g}{N}$ for small fixed g.

Consider a single guiding conception in this network, and denote its elements 1 and 2. Each element has a number of children for which it is first parent, and a number for which it is second parent. Let h_1 be the number of children element 1 has for which it is first parent, and k_1 be the number of children for which it is second parent. Similarly let h_2 be the number of children element 2 has for which it is first parent, and k_2 the number of children for which it is second parent. The random variables h_1 and h_2 have a joint distribution given by a multinomial distribution: there are mN bottom elements, and of these h_1 connect with top element 1, h_2 with top element 2, and the remainder with the other $N-2$ top elements:[1]

$$Prob(h_1, h_2) = \binom{mN}{h_1, h_2}(\frac{1}{N})^{h_1}[\frac{1}{N}]^{h_2}[1-\frac{2}{N}]^{[mN-h_1-h_2]}$$

Note that according to this model h_1 and h_2 are jointly distributed since any given bottom element can only have a single first parent. However, for large-scale random networks based on the given model this jointness is very small. In principle h_1 has range 0 to mN, and if h_1 has value j, the range of h_2 is from 0 to $mN-j$. However, for a large-scale random network based on the given model each top element will have only a few first children. Thus h_1 and h_2 each effectively range from 0 to some modest positive number far below mN. This has two implications. First, we can take the maximum limit for the number of first parent children as far below mN. Second, because the range of h_2 is thus fixed and does not depend on h_1, the dependence between h_1 and h_2 is broden and h_1 and h_2 can be viewed as approximately independent which simplifies the calculation. Indeed for this case the expected values of h_1 and h_2 are each m and thus the expected size of this part of the seed pool is m^2.

Now consider the adjustments to this expected value. First, a bottom element that has element 1 as first parent can only be part of the seed pool if it does not connect with element 2 as a second parent, and this probability is $1-p$; likewise for elements having 2 as first parent. This applies to each element, hence the expected size of the pool is reduced by the factor $(1-p)^2$. Second, if any pair of elements share a second parent, they are ruled out of the pool. The probability they share any given parent is p^2, and there are $N-2$ remaining top elements; hence the overall probability that a given pair does not share a second parent is $(1-p^2)^{N-2}$. Due to the linearity of the expectation, this term applies to each pair in the pool independently.[2] Now consider the case in which p is on order $\frac{1}{N}$ or $\frac{g}{N}$. The first correction is approximately $1-\frac{2g}{N}$ which is very close to 1 for large N and small g. For the second term, rewrite it as $exp^{(N-2)log(1-p^2)}$ and applying the expansion $log(1-x)$ is approximately $-x$ yields a value for this of approximately $exp^{-\frac{g^2}{N}}$ which for large N and small g is negligibly different from 1. Thus the expected first parent seed pool size is within $2g/N$ of m^2.

[1] For calculations over multinomial distributions used see for example: http://utstat.toronto.edu/brunner/oldclass/312f12/lectures/312f12MultinomialHandout.pdf .

[2] Pick one of the first-parent children of element 1, say x. It may have 1, 2,... second parents. Conditional on it having j second parents, it's probability of not overlapping any specific first-parent child of element 2 is simply $(1-p)^j$ and this value is independent across the children of 2 thus simply $h_2(1-p)^j$ times the probability x has j children r_j. Now simply sum over j and regardless of the values of r_j the overall sum will be h_2 times the probability that x does not intersect with any given child of element 2, thus linearly additive. I compute this last probability directly in the text.

Conditional on (h_1, h_2), the random variables k_1 and k_2 are independent; each is a binomial distribution based on the number of remaining bottom elements - for k_1 its binomial has maximum value $mN - h_1$, and for k_2 its binomial has maximum value $mN - h_2$. However, if, again, h_1 and h_2 are small relative to mN, and given p on the order of $\frac{1}{N}$ so that the k_1 and k_2 distributions will concentrate on small values, these exact maximum values are not relevant. Again, this has the implication that k_1 and k_2 are approximately independent for the kinds of networks under consideration without needing to condition on h_1 and h_2; and each follows a binomial distribution for which the mass is concentrated at small values. The expected values of k_1 and k_2 are $p(mN - h_1)$ and $p(mN - h_2)$; for p small and on the order of $\frac{1}{N}$ or equal to $\frac{g}{N}$ both are approximately mg. So the product is m^2g^2. Now for a member of the k_1 set to be a valid partner it must not link with element 2either as a first or second parent: the probability of this is $\frac{N-1}{N}(1-p)$. Thus the size of the (k_1, k_2) part of the seed pool is in expectation $m^2g^2\frac{(N-1)(N-g)}{N^2}^2$. For N large and g small this last correction is within $\frac{(1+g)}{N}$ of 1. Finally, for each pair of elements from the k_1 and k_2 sets they are a valid pair only if they do not share either a first parent or any second parent. The calculation is similar to the calculation above for (h_1, h_2) pairs, with the one difference that the possibility of an overlap as first parents or first parent-second parent must also be taken into account. Consider such a pair. The probability their two first parents are the same is $\frac{1}{N-2}$ and the probability they are different is thus $1 - \frac{1}{N-2}$. Given different first parents, the probability that the k_2 element does not have a second parent equal to the first parent for k_1 is just $1 - p$, and likewise for k_1 with k_2's first parent. Finally, conditional on not sharing first parents or first with second parents, the probability they do not share any second parent in common is $(1-p^2)^{N-4}$ where the 4 subtracted from N represents the two guiding conception elements and the two first parents of k_1 and k_2 (which must be different). Thus the overall probability the pair is valid is $[1 - \frac{1}{N-2}](1-p)^2(1-p^2)^{N-4}$. Setting p equal to $\frac{g}{N}$, the very last term is negligibly different from 1 and the first terms are $\frac{(N-3)(N-g)^2}{(N-2)N^2}$, on the order of $1 - \frac{2g}{N}$. Thus again this correction is also small and the expected seed pool size is within $\frac{2g}{N}$ of m^2g^2.

The two additional parts to the pool: (h_1, k_2) and (h_2, k_1) mixed first parent - second parent pairs. Consider the pairing of an element in the h_1 pool with an element in the k_2 pool. Given that k_2 does not have element 1 as its first parent, to compute the probability they form a valid pair we need to compute the probability k_2's first parent is not a second parent to h_1, and that they do not share any second parents. The probability of the first event is simply $1 - p$ and the probability of the second is $(1-p^2)^{N-3}$ where we subtract 3 from N for the two guiding conception elements and k_2's first parent; so the probability they form a valid pair is the multiplication of these two terms, and the familiar argument shows that this is within $\frac{g}{N}$ of 1. Thus the expectation of this part of the seed pool is within this tolerance of m^2g. The same hold for the h_2-k_1 part of the pool.

Thus the argument overall shows that the expected seed pool size in this kind of large-scale, sparse random network is within a linear term in $p = \frac{g}{N}$ of:

$$m^2 + m^2g^2 + 2m^2g = m^2(1+g)^2.$$

Random Search - Expected Number of Trials to Find First of Three Seed Pairs

The N pairs form a sequence of integers from 1 to N. The first of these pairs can be in any position from 1 to $N - 2$; the remaining two of the pairs are after the first in the sequence (this is why the first cannot be any later than $N - 2$). As noted in the main text this does not take into account the constraint that there cannot be two golden seed pairs under the same pair of top element parents. Taking this into account requires also knowing the number of children in each pool; if this is constant as for the symmetric hierarchy this will reduce N by a number equal to the pool size times two, since the first golden seed pair must be found

by the third from the last pool. This requires a more sophisticated search approach than strictly random and I do not consider it in this section.

Consider now searching over the $N-2$ pairs. The likelihood of any position m in this range is $\frac{1}{N-2}$. Once the first pair is placed, the next two must be positioned after it, and the probability of this is $\left[\frac{N-m}{N-1}\right]\left[\frac{N-m-1}{N-2}\right]$. Lastly, if the first is in position m than it requires m trials to discover it. Thus the overall formula for the expected number of trials is:

$$\binom{3}{1}\frac{1}{(N-1)(N-2)^2}\sum_{m=1}^{N-2} m(N-m)(N-m-1)$$

where all terms in denominators have been moved outside the sum. The sum has three distinct terms: (i) $m(N)(N-1)$; (ii) m^3; and (iii) $-m^2(2N-1)$. Each of these is a power in m times a constant, and the sum runs from 1 to $N-2$. Thus formulas for the sum of integers over a range, the sum of squares of integers over the range, and the sum of cubes of integers over the range apply (see https://brilliant.org/wiki/sum-of-n-n2-or-n3/). These 3 formulas in this case are:

$$\sum_{m=1}^{N-2} m = \frac{(N-2)(N-1)}{2}$$

$$\sum_{m=1}^{N-2} m^2 = \frac{(N-2)(N-1)(2N-3)}{6}$$

$$\sum_{m=1}^{N-2} m^3 = \frac{(N-2)^2(N-1)^2}{4}$$

The expected number of trials for any value of N can then be computed. When N is large we can assume $N-1$ and $N-2$ are approximately the same as N, and $2N-3$ is approximately the same as $2N$. With these simplifications the overall formula from above becomes:

$$\frac{3}{N^3}\left[\frac{N^4}{2}+\frac{N^4}{4}-\frac{4N^4}{6}\right]$$

which is:

$$3N\left[\frac{1}{2}+\frac{1}{4}-\frac{2}{3}\right]$$

or $\frac{N}{4}$.

Distribution for Number of Guiding Conceptions that Cover a Given Seed Pair

I will show first how to compute the distribution associated with the number of guiding conceptions that covers a given seed pair beyond the first parent pair. There are $N-2$ top level elements that are not first parent to either seed element. For each of these there are these possibilities: (i) one of the seed elements links to this top element and not the other - there are two ways this can happen, each has probability $p(1-p)$; or (ii) neither seed element links, probability $(1-p)^2$. Note that it cannot be the case that both seed elements link to this top element since in that case they would share a common parent and would not be a valid seed. Hence applying conditional probability, the probability one seed element links and not the other is $q=\frac{p(1-p)}{2p(1-p)+(1-p)^2}$. Since links are independent, we can use the multinomial distribution to compute the probability that, out of the $N-2$ relevant top elements, g_1 link to the first seed element, and g_2 to the second:

$$\binom{N-2}{g_1,g_2}\, q^{(g_1+g_2)}\,[1-2q]^{[(N-2)-g_1-g_2]}$$

Given g_1 and g_2, it follows that there are $(1+g_1)(1+g_2)$ total guiding conceptions that cover this seed pair, which is the total number of ways of combining the top elements that cover each seed element. The distribution is then evaluated over all pairs g_1, g_2, with each $g_i \geq 0$ and $g_1+g_2 \leq M_H-2$.