Stochastic Dynamics of Lexicon Learning in an Uncertain and Nonuniform World

Rainer Reisenauer

Physik-Department, Technische Universität München, James-Franck-Strasse 1, 85748 Garching, Germany and SUPA, School of Physics and Astronomy, University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom

Kenny Smith

Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh EH8 9AD, United Kingdom

Richard A. Blythe

SUPA, School of Physics and Astronomy, University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom

(Received 22 February 2013; published 21 June 2013)

We study the time taken by a language learner to correctly identify the meaning of all words in a lexicon under conditions where many plausible meanings can be inferred whenever a word is uttered. We show that the most basic form of cross-situational learning—whereby information from multiple episodes is combined to eliminate incorrect meanings—can perform badly when words are learned independently and meanings are drawn from a nonuniform distribution. If learners further assume that no two words share a common meaning, we find a phase transition between a maximally efficient learning regime, where the learning time is reduced to the shortest it can possibly be, and a partially efficient regime where incorrect candidate meanings for words persist at late times. We obtain exact results for the word-learning process through an equivalence to a statistical mechanical problem of enumerating loops in the space of word-meaning mappings.

DOI: 10.1103/PhysRevLett.110.258701 PACS numbers: 01.40.Ha, 05.70.Fh, 87.19.lv, 89.65.Ef

On average, children learn ten words a day, thereby amassing a lexicon of 60 000 words by adulthood [1]. This speed of learning is remarkable given that every time a speaker says a word, a hearer cannot be certain of its intended meaning [2]. Our aim is to identify which of the many proposed mechanisms for eliminating uncertainty can actually deliver such rapid word learning. In this Letter, we pursue this aim within the long tradition of applying quantitative methods from statistical mechanics to problems in learning [3–6] and communication [7–9].

Empirical research suggests that two basic types of learning mechanisms are involved in word learning. First, a learner can apply various heuristics—e.g., attention to gaze direction [10] or prior experience of language structure [11]—at the moment a word is produced to hypothesize a set of plausible meanings. However, these heuristics may leave some residual uncertainty as to a word's intended meaning in a single instance of use. If the heuristics are weak, the set of candidate meanings could be very large. This residual uncertainty can be eliminated by comparing separate instances of a word's use: if only one meaning is plausible across all such instances, it is a very strong candidate for the word's intended meaning. This second mechanism is referred to as cross-situational learning [12,13]. Formally, it can be couched as a process whereby associations between words and meanings are strengthened when they co-occur [13–16], as in neural network models for learning [3–6,17]. It can also be viewed as an error-correction process [7–9] where a target set of associations is reconstructed from noisy data.

There is little consensus as to which word-learning mechanisms are most important in a real-world setting [18–22]. In part this is because word-learning experiments (e.g., Refs. [20,23,24]) are necessarily confined to small lexicons. A major question is whether strategies observed in experiments allow realistically large lexicons to be learned rapidly: this can be fruitfully addressed through stochastic dynamical models of word learning [15,25–27]. In these models, a key control parameter is the context size: the number of plausible, but unintended, meanings that typically accompany a single word's true meaning. Even when contexts are large, the rapid rate of learning seen in children is reproduced in models where words are learned independently by cross-situational learning [15,25–27]. This suggests that powerful heuristics capable of filtering out large numbers of spurious meanings are not required. However, a recent simulation study [28] shows that this conclusion relies on the assumption that these unintended meanings are uniformly distributed. In the more realistic scenario where different meanings are inferred with different probabilities, word-learning rates can decrease dramatically as context sizes increase. Powerful heuristics may be necessary after all.

One heuristic of great interest to empiricists (e.g., Refs. [29–32]) and modelers (e.g., Refs. [15,26–28,33]) is a mutual exclusivity constraint [29]. Here, a learner

assumes that no two words may have the same meaning. This generates nontrivial interactions between words which makes analysis of the corresponding models difficult. For example, if one begins with a master equation, as in Refs. [15,25,26], the expressions become unwieldy to write down, let alone solve. Here, we adopt a fundamentally different approach which entails identifying the criteria that must be satisfied for a lexicon to be learned. This allows the existing results for the simple case of independently learned words and uniform meaning distributions [26] to be generalized to arbitrary meaning distributions and exactly solves the interacting problem to boot. Our main result is that mutual exclusivity induces a dynamical phase transition at a critical context size, below which the lexicon is learned at the fastest possible rate (i.e., the time needed to encounter each word once). As far as we are aware, the ability of a single heuristic to deliver such fast learning has not been anticipated in earlier work.

We begin by defining our model for lexicon learning. The lexicon comprises W words, and each word i is uttered as a Poisson process with rate ϕ_i. In all cases, we take words to be produced according to the Zipf distribution $\phi_i = 1/(\mu i)$ that applies for the $\sim 10^4$ most frequent words in English [34–36]. Here, $\mu = \sum_{i=1}^{W}(1/i)$ so that one word appears on average per unit time. Each time a word i is presented, the intended target meaning is assumed always to be inferred by the learner by applying some heuristics. At the same time, a set of nontarget confounding meanings called the context is also inferred.

In the purest version of cross-situational learning [13,26], a learner assumes that all meanings that have appeared every time a word has been uttered are plausible candidate meanings for that word. The word becomes learned when the target is the only meaning to have appeared in each episode. In the noninteracting case, each word is learned independently—see Fig. 1(a). In the interacting case, mutual exclusivity acts to further exclude the meanings of learned words as candidates for other words. We take this exclusion to occur at the instant a word is learned, which means a single learning event may trigger an avalanche of other learning events by repeated application of mutual exclusivity. An example of this nontrivial effect that is hard to handle within standard approaches [15,26] is shown in Fig. 1(b). Here, learning "square" causes "circle" to be learned at the same time.

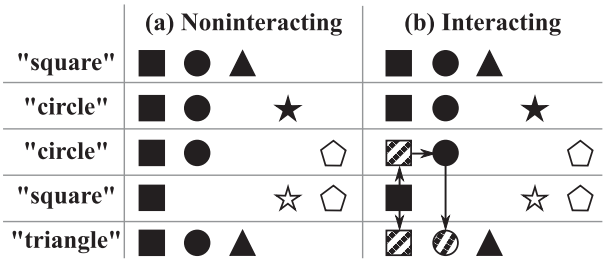

FIG. 1. Acquisition of a three-word lexicon. Solid shapes are meanings that have appeared in every episode alongside a word; open shapes are therefore excluded as candidate meanings. (a) In the noninteracting case, only the meaning of the word "square" is learned. (b) In the interacting case, mutual exclusivity further removes meanings (shown hatched) of learned words, both prospectively and retrospectively (shown by arrows). All three words are learned in this example.

We consider the noninteracting case first both to introduce our more powerful analytical approach and to pinpoint the origin of the catastrophic increase in learning times noted in Ref. [28]. Two conditions must be satisfied for the lexicon to be learned by a given time: (C1) all words must have been exposed at least once, and (C2) no confounding meaning may have appeared in every episode that any given word was uttered. To express these conditions mathematically, we introduce two stochastic indicator variables. We take $E_i(t) = 1$ if word i has been uttered before time t, and zero otherwise, and $A_{i,j}(t) = 1$ if confounding meaning j has appeared in every context alongside word i up to time t (or if word i has never been presented), and zero otherwise. Conditions (C1) and (C2) then imply that the probability that the lexicon has been learned by time t is

$$L(t) = \left\langle \prod_i E_i(t) \prod_{j\neq i} [1 - A_{i,j}(t)] \right\rangle = \left\langle \prod_{i\neq j} [1 - A_{i,j}(t)] \right\rangle, \quad (1)$$

where the angle brackets denote an average over all sequences of episodes that may occur up to time t. The second equality holds because $A_{i,j}(t) = 1\ \forall j \neq i$ if $E_i(t) = 0$.

This expression is valid for any distribution over contexts. For brevity, we consider a single, highly illustrative construction that we call resampled Zipf (RZ). It is based on the idea that meaning frequencies should follow a similar distribution to word forms [28]. It works by associating an ordered set $\mathcal{M}_i$ of M confounding meanings with each word i. The kth meaning in each set has an *a priori* statistical weight $1/k$. Whenever word i appears, meanings are repeatedly sampled from $\mathcal{M}_i$ with their *a priori* weights and added to the context if they are not already present until a context of C distinct meanings has been constructed. When words are learned independently, the learning time depends only on M, W, and C, and not on which meanings are present in any given set $\mathcal{M}_i$ [26].

We seek the time t^* at which the lexicon is learned with some high probability $1 - \epsilon$. In the RZ model, each context is an independent sample from a fixed distribution. Hence, the correlation functions $\langle A_{i_1,j_1} A_{i_2,j_2} \ldots \rangle$ in Eq. (1) all decay exponentially in time. To find t^* to good accuracy in the small-ϵ limit, only the slowest decay mode for each word i is needed. Higher-order correlation functions depend on many meanings co-occurring and decay more rapidly than lower-order correlation functions. Thus, at late times, Eq. (1) is well approximated by (see Ref. [37])

$$L(t) \sim \prod_i [1 - e^{-\phi_i(1-a_i^*)t}], \tag{2}$$

where a_i^* is the fraction of episodes in which word i's most frequent confounder appears alongside the target. This expression generalizes results for independently learned words [15,25,26] from uniform to arbitrary nonuniform confounder distributions.

The RZ model has the further simplification that a_i^* has a common value a^* for all words i. Then, it is known from previous calculations [26] for Zipf-distributed word frequencies that the learning time is

$$t^* \sim \frac{\mu W}{1 - a^*} \mathcal{W}_0\left(\frac{W}{-\ln(1-\epsilon)}\right), \tag{3}$$

where $\mathcal{W}_0(z)$ is the principal branch of the Lambert W function [38]. For large argument, this function behaves as a logarithm.

In Fig. 2, we compare the analytical result [Eq. (3)] with learning times obtained from direct Monte Carlo simulations conducted as detailed in Ref. [26]. The only complication is that we unfortunately have no analytic expression for a^* arising from the RZ procedure. We therefore obtain the frequency of the most common confounder for given C and M from independent Monte Carlo samples. The agreement between Eq. (3) and simulation is very good.

Figure 2 also shows that the learning time increases superexponentially with the context size. We have found that the probability the kth most confounder appears in a context of size C fits the form $p_k \approx 1 - (1 - w_k)^{Ce^{\lambda C}}$ where w_k is the *a priori* probability and λ is a fitting parameter that depends on M and k. As noted by Vogt [28], the repeated sampling without replacement implies that $p_k \geq 1 - (1 - w_k)^C$. Our analysis further reveals that the learning time is entirely determined by the frequency of

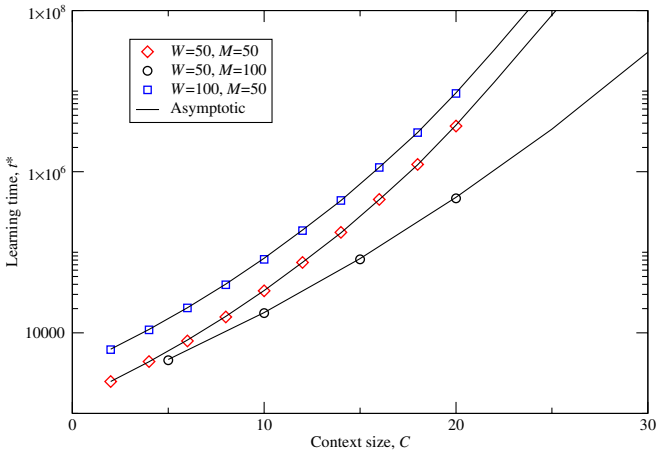

FIG. 2 (color online). Time to learn a lexicon of W words independently to a residual probability $\epsilon = 0.01$ with C of M confounders present in each episode. Points: data from Monte Carlo simulations (over 10 000 sampled lexicons in each case). Lines: the analytical result, Eq. (3).

the most common confounder a^* through Eq. (3). We note that this is true even when other confounders have comparable appearance frequencies ($C \leq 5$).

We now turn to the case where the mutual exclusivity constraint serves to exclude the meanings of learned words as possible meanings for other words. In this case, it is important to distinguish between labeled and unlabeled meanings: an unlabeled meaning is not the target meaning of any word in the lexicon and hence cannot be excluded using the mutual exclusivity constraint. To generalize Eq. (1) to this problem, we must identify the conditions for the lexicon to be learned. Condition (C1) still applies: each word must be uttered at least once for a learner to be able to learn it. Condition (C2) now applies only to unlabeled confounding meanings: these can only be excluded if they fail to appear in a context, as before. When these two conditions are satisfied, there is a third—necessary and sufficient—condition for the lexicon to be learned that takes into account all the interactions and avalanches generated by the mutual exclusivity constraint. This is condition (C3): no candidate loops exist at time t. A candidate loop, $\ell = (i_1, i_2, \ldots, i_n)$ is a subset of distinct, labeled meanings whereby each meaning i_k has appeared alongside the word associated with meaning i_{k-1} (or i_n if $k = 1$) every time it has been uttered. Inspection of Fig. 1(b) shows that the one candidate loop (filled square, filled circle) that exists after the third episode is destroyed in the fourth. Then, in the fifth episode, the final word appears, and since no unlabeled meaning is a candidate for any word, the entire three-word lexicon is learned.

To see why condition (C3) is necessary and sufficient in general when (C1) and (C2) hold, we first show that a candidate loop must exist if the lexicon has not been learned. Suppose word i_1 has not been learned. Then, at least one meaning, i_2, must confound word i_1. Word i_2 must also not have been learned, otherwise meaning i_2 would not confound word i_1. Hence, word i_2 must be confounded by a meaning i_3, and so on. As there is a finite set of words, this sequence of meanings must eventually form a loop.

We now show the lexicon cannot have been learned if a candidate loop exists by first assuming that it has been learned under these conditions. Then, if word i_1 was learned at time t, word i_2 must have been learned before time t for mutual exclusivity to act (even if words i_1 and i_2 are learned as part of the same avalanche). Iterating this argument around the loop, one finds that word i_1 can only have become learned at time t if it had already been learned at some earlier time. This contradiction therefore implies that the absence of candidate loops and a learned lexicon are equivalent.

We again use indicator variables to translate conditions (C1)–(C3) into an exact expression for the learning probability. Introducing $C_\ell(t) = A_{i_1,i_2}(t)A_{i_2,i_3}(t)\cdots A_{i_n,i_1}(t)$ that equals 1 if the loop ℓ persists at time t, we have

$$L(t) = \left\langle \prod_{i=1}^{W} E_i(t) \prod_{j>W} [1 - A_{i,j}(t)] \prod_{\ell} [1 - C_\ell(t)] \right\rangle, \quad (4)$$

again valid for any distribution of confounding meanings. Here, meanings 1 to W correspond to words 1 to M, and so meanings with an index $j > W$ are unlabeled. The product over ℓ is over all possible candidate loops. This expression has the remarkable property that it is expressed concisely in terms of the word and confounder appearance frequencies alone: the avalanche dynamics triggered by mutual exclusivity do not enter explicitly. This property, reminiscent of the avalanche dynamics of Abelian sandpile models [39], reduces analysis of the learning probability to the statistical mechanical problem of enumerating candidate loops.

In the interacting problem, the structure of each candidate set $\mathcal{M}_i$ is important, as this determines which words interact. We consider a model which has no unlabeled meanings and where each set $\mathcal{M}_i$ is a sample of M nontarget meanings obtained via the RZ prescription. Then, in each episode, C meanings are drawn from the relevant candidate set using RZ again, but with an *a priori* weight $1/k$ where k is the rank of a meaning within the set $\mathcal{M}_i$ when ordered by the frequency of the corresponding words. Thus, meanings of high-frequency words are high-frequency confounders. Learning times from Monte Carlo simulations are shown in Fig. 3.

We observe two distinct learning-time regimes. At small C, the learning time is constant, and close to the time it takes for all words in the lexicon to appear at least once. [This time is given by Eq. (3) with $a^* = 0$.] In this regime, learning is as fast as it can possibly be: mutual exclusivity is maximally efficient and reverses the undesirable increase in learning times that arises from nonuniform confounder distributions. Above a critical context size, the learning time rises but remains much smaller than when words are learned independently: mutual exclusivity is partially efficient in this regime.

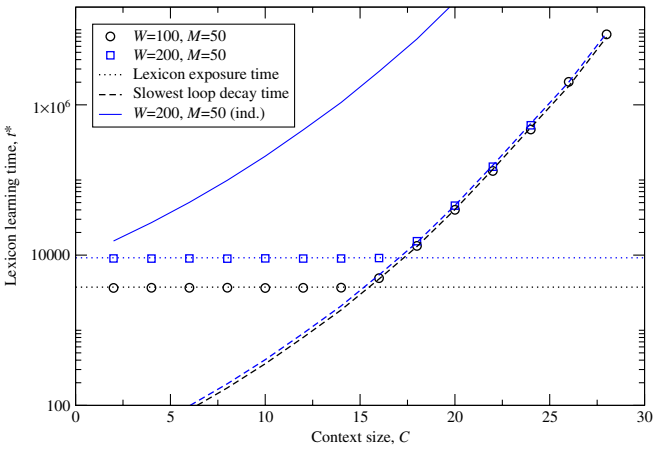

FIG. 3 (color online). As Fig. 2 but with the mutual exclusivity constraint. Points: data from Monte Carlo simulations (100 000 lexicons for $C \leq 20$, at least 2500 lexicons for larger C). Dotted lines: time for the entire lexicon to have been exposed with residual probability $\epsilon = 0.01$. Dashed lines: time for the slowest decaying candidate loop to remain with probability ϵ. Solid line: time to learn lexicon independently, Eq. (3), for comparison.

Our exact result [Eq. (4)] can be used to explain these observations [40]. For the RZ model as described above, it turns out that only one confounder loop $\ell = (1, 2)$ is relevant at late times. Consequently, the learning probability $L(t)$ is asymptotically given as the product of two factors. The first gives the probability that all words have been encountered by time t and approaches unity exponentially with rate $1/\mu W$. The second is the probability that the loop $\ell = (1, 2)$ has not decayed away: this approaches unity with rate $3(1 - a^*)/2\mu$. The appearance frequency of the most frequent confounder a^* increases with context size. When $a^* < 1 - (2/3W)$, the slowest relaxational mode of the learning probability is associated with each word being uttered at least once, whereas for larger values, the slowest mode comes from eliminating the confounder loop. In this latter partially efficient regime, the lexicon learning time is predicted as $t^* = -(2\mu \ln\epsilon/3(1 - a^*)$ for small ϵ, in very good agreement with simulation data (see Fig. 3). We describe the sudden change in the dominant relaxational behavior—a phenomenon seen also in driven diffusive systems [41]—as a dynamical phase transition. It is broadly reminiscent of transitions exhibited by combinatorial optimization problems, whereby the number of unsatisfied constraints increases from zero above a critical difficulty threshold [42]. In the present case, the learning problem remains solvable in both regimes, but there is a transition from a regime where it is solved in constant time to one where the time grows superexponentially in the difficulty of the problem (here, the context size).

To summarize, we have found that mutual exclusivity is an extremely powerful word-learning heuristic. It can yield lexicon learning times in the presence of uncertainty that coincide with the time taken for each word to be heard at least once. Empirical data (summarized in Ref. [26]) suggest that this is easily fast enough for realistic lexicons of $W = 60\,000$ words to be learned. To enter the partially efficient regime, each word's most frequent confounder would need to be present in at least 99.99% of all episodes: even then, learning is over W times faster than when mutual exclusivity is not applied. The dynamical transition between a maximally and partially efficient regime also appears to be present in a variety of word-learning models we have investigated, e.g., those in which confounder frequencies are uncorrelated with their corresponding word frequencies or using less memory-intensive learning strategies [43]. We also expect the transition to be evident in models where the target meaning does not always appear, at least in the regime where learning is possible [15,27]. We believe the analytical methods introduced in

this Letter should allow more detailed quantities to be calculated, e.g., the distribution of learning times for a given word, which would shed light on such phenomena as the childhood vocabulary explosion at around 18 months [44]. Similar thinking may also allow analysis of other nonequilibrium dynamical systems whose master equations are hard to solve directly. Finally, our results suggest new empirical questions, such as whether high-frequency confounders correlate with high-frequency words, and the extent to which learners are able to apply the mutual-exclusivity constraint retroactively. We therefore contend that statistical physicists can contribute much to the understanding of how children learn the meaning of words.

We thank Mike Cates and Cait MacPhee for comments on the manuscript.

[1] P. Bloom, *How Children Learn the Meanings of Words* (MIT Press, Cambridge, MA, 2000).
[2] W. V. O. Quine, *Word and Object* (MIT Press, Cambridge, MA, 1960), pp. 26–79.
[3] T. L. H. Watkin and A. Rau, Rev. Mod. Phys. **65**, 499 (1993).
[4] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[5] D. J. Amit, H. Gutfreund, and H. Sompolinksy, Phys. Rev. A **32**, 1007 (1985).
[6] D. J. Amit, H. Gutfreund, and H. Sompolinksy, Phys. Rev. Lett. **55**, 1530 (1985).
[7] N. Sourlas, Nature (London) **339**, 693 (1989).
[8] Y. Kabashima, T. Murayama, and D. Saad, Phys. Rev. Lett. **84**, 1355 (2000).
[9] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, Oxford, 2001).
[10] D. A. Baldwin, Child Dev. **62**, 875 (1991).
[11] P. Bloom and L. Markson, Trends Cogn. Sci. **2**, 67 (1998).
[12] S. Pinker, *Learnability and Cognition: The Acquisition of Argument Structure* (MIT Press, Cambridge, MA, 1989).
[13] J. M. Siskind, Cognition **61**, 39 (1996).
[14] P. Vogt and A. D. M. Smith, in *Proceedings of the Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn)*, edited by T. Lenaerts, A. Nowe, and K. Steenhout (Vrije Universiteit Brussel, Brussels, 2004).
[15] P. F. C. Tilles and J. F. Fontanari, J. Math. Psychol. **56**, 396 (2012).
[16] C. Yu and L. B. Smith, Psychol. Rev. **119**, 21 (2012).
[17] F. Pulvermüller, Behav. Brain Sci. **22**, 253 (1999).
[18] L. Gleitman, Lang. Acquis. **1**, 3 (1990).
[19] S. Pinker, Lingua **92**, 377 (1994).
[20] C. Yu and L. B. Smith, Psychol. Sci. **18**, 414 (2007).
[21] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, Psychol. Sci. **20**, 578 (2009).
[22] T. N. Medina, J. Snedeker, J. C. Trueswell, and L. R. Gleitman, Proc. Natl. Acad. Sci. U.S.A. **108**, 9014 (2011).
[23] L. Smith and C. Yu, Cognition **106**, 1558 (2008).
[24] K. Smith, A. D. M. Smith, and R. A. Blythe, Cogn. Sci. **35**, 480 (2011).
[25] K. Smith, A. D. M. Smith, R. A. Blythe, and P. Vogt, in *Symbol Grounding and Beyond*, Lecture Notes in Computer Science Vol. 4221 (Springer, Heidelberg, 2006), pp. 31–44.
[26] R. A. Blythe, K. Smith, and A. D. M. Smith, Cogn. Sci. **34**, 620 (2010).
[27] P. F. C. Tilles and J. F. Fontanari, Europhys. Lett. **99**, 60 001 (2012).
[28] P. Vogt, Cogn. Sci. **36**, 726 (2012).
[29] E. M. Markman and G. F. Wachtel, Cogn. Psychol. **20**, 121 (1988).
[30] R. M. Golinkoff, C. B. Mervis, and K. Hirsh-Pasek, J. Child Lang. **21**, 125 (1994).
[31] J. Halberda, Cognition **87**, B23 (2003).
[32] E. M. Markman, J. L. Wasow, and M. B. Hansen, Cogn. Psychol. **47**, 241 (2003).
[33] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, Adv. Neural Inf. Process. Syst. **20**, 20 (2007).
[34] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Cambridge, MA, 1949).
[35] R. Ferrer i Cancho and R. V. Solé, J. Quant. Ling. **8**, 165 (2001).
[36] A. M. Petersen, J. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, Sci. Rep. **2**, 313 (2012).
[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.110.258701 for a step-by-step derivation.
[38] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, Adv. Comput. Math. **5**, 329 (1996).
[39] D. Dhar, Phys. Rev. Lett. **64**, 1613 (1990).
[40] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.110.258701 for details.
[41] J. de Gier and F. H. L. Essler, Phys. Rev. Lett. **95**, 240601 (2005).
[42] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).
[43] R. Reisenauer, K. Smith, A. D. M. Smith, and R. A. Blythe (to be published).
[44] B. McMurray, Science **317**, 631 (2007).