Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Phase transitions in least-effort communications

Mikhail Prokopenko[1,2], **Nihat Ay**[1,3], **Oliver Obst**[2] **and Daniel Polani**[4]

[1] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22-26, Leipzig 04103, Germany
[2] CSIRO Information and Communications Technologies Centre, PO Box 76, Epping, NSW 1710, Australia
[3] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
[4] Department of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK
E-mail: prokopenko.mikhail@gmail.com, nay@mis.mpg.de, oliver.obst@csiro.au and d.polani@herts.ac.uk

Received 14 July 2010
Accepted 21 October 2010
Published 15 November 2010

Online at stacks.iop.org/JSTAT/2010/P11025
doi:10.1088/1742-5468/2010/11/P11025

Abstract. We critically examine a model that attempts to explain the emergence of power laws (e.g., Zipf's law) in human language. The model is based on the principle of least effort in communications—specifically, the overall effort is balanced between the speaker effort and listener effort, with some trade-off. It has been shown that an information-theoretic interpretation of this principle is sufficiently rich to explain the emergence of Zipf's law in the vicinity of the transition between referentially useless systems (one signal for all referable objects) and indexical reference systems (one signal per object). The phase transition is defined in the space of communication accuracy (information content) expressed in terms of the trade-off parameter. Our study explicitly solves the continuous optimization problem, subsuming a recent, more specific result obtained within a discrete space. The obtained results contrast Zipf's law found by heuristic search (that attained only local minima) in the vicinity of the transition between referentially useless systems and indexical reference systems, with an inverse-factorial (sub-logarithmic) law found at the transition that corresponds to global minima. The inverse-factorial law is observed to be the most representative frequency distribution among optimal solutions.

Keywords: exact results, stochastic search, communication, supply and information networks

1742-5468/10/P11025+30$30.00

Contents

1. Introduction

To put it simply, Zipf's law states that given some (natural language) text and the ranking of its words in the order of decreasing frequency, the frequency of any word is inversely proportional to its rank. Thus, within such *frequency distributions* of words sorted by decreasing frequencies, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.

Zipf's law has been reliably observed in data from multiple diverse textual sources, while it has been shown that random texts do not exhibit Zipf's law-like rank distribution [1]. At the same time, widely accepted theoretical explanations for Zipf's law are still lacking. As mentioned by Manin [2], 'Zipf's law (1949) may be one of the most enigmatic and controversial regularities known in linguistics. It has been

J. Stat. Mech. (2010) P11025

alternatively billed as the hallmark of complex systems and dismissed as a mere artefact of data presentation. The simplicity of its formulation, its experimental universality, and its robustness starkly contrast with the obscurity of its meaning'. According to Ferrer i Cancho [3], various subsets of the language (e.g., subsamples consisting of nouns only in multi-author collections of texts, the speech of schizophrenics and very young children, military communications) obey the generalized Zipf's law, that is, follow a power law $P(\rho) \propto \rho^{-\alpha}$, where ρ is the rank, $P(\rho)$ is the frequency of the word having rank ρ, and the exponent α may differ from 1.

Furthermore, it is well known now that power-law distributions occur in a diverse range of physical, biological, technological and social phenomena [4, 5]. A power-law distribution is often called a scale-free distribution—it satisfies the property that $P(bx) = g(b)P(x)$, for any b, and some function g that depends on the exponent of the power law. That is, if the scale of units by which x is measured is increased by a factor of b, the shape of the distribution $P(x)$ is unchanged, except for some multiplicative constant [4].

The widespread universality of Zipf's law and power laws in general has generated many attempts at an explanation [5]. In this study we critically examine a model, proposed by Ferrer i Cancho and Solé [6] and comprehensively expanded by Ferrer i Cancho [7], that attempts to explain the emergence of power laws (e.g., Zipf's law) in human language as a result of minimizing a communication effort that balances certain trade-offs. The original model [6] used an assumption that the objects referred to in the communication system are uniformly distributed (the uniformity assumption), while the expanded model [7] relaxed this assumption, using a more general energy (cost) function.

These models [6, 7] formalized the principle of least effort as an optimization problem, and suggested a candidate mechanism for generating power laws—by heuristically solving the optimization problem and considering a resultant phase transition. Another follow-up study by Ferrer i Cancho and Díaz-Guilera [8] argued that the optimal solutions found by this method attain, however, only local minima. Ferrer i Cancho and Díaz-Guilera analytically derived global minima of the cost Ω_λ (for a discrete system), showing that the phase transition is in fact a step function. They also discussed the difficulties of explaining power laws observed in natural languages using the proposed models.

The study presented here contrasts the results of Ferrer i Cancho and Solé [6] (obtained only for local minima) with explicit global solutions of the *continuous* optimization problem. Using a specific characterization of minimal solutions of the continuous optimization problem (representable as suitably defined functions), we confirm that (i) the phase transition is a step function, and (ii) the minimal solutions have no synonyms, generalizing observations by Ferrer i Cancho and Díaz-Guilera [8] for a discrete system. This leads to the conclusion that power laws are not a necessary consequence of such optimization.

Additionally, we derive a necessary condition required for the emergence of power laws within communication systems. This condition places an extra constraint on the involved communication efforts.

The presented results point to a sub-logarithmic dependency as the most representative frequency distribution among optimal solutions, rather than a power law. Specifically, we show that the model [6] is not strong enough to produce power laws at the global minima, where instead another dependency is shown to be more dominant

J. Stat. Mech. (2010) P11025

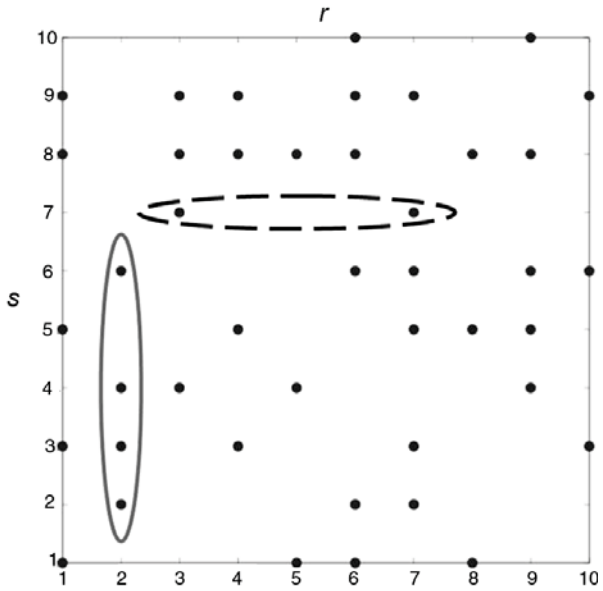

Figure 1. The binary matrix **A**, with dots indicating non-zero elements. Synonyms are enclosed within the vertical solid oval. Polysemy is shown with the horizontal dashed-line oval.

(for systems with an equal number of signals and objects): an inverse-factorial (sub-logarithmic) law.

1.1. Basic information-theoretic model

The model introduced by Ferrer i Cancho and Solé [6] provided an information-theoretic framework for the principle of least-effort communications, and in addition put forward a candidate mechanism for generating power laws in communication systems. The latter aspect is particularly important, given the current debate on the origin of power laws. For instance, Kosmidis *et al* [9] relate their statistical mechanical approach to human language to 'the pioneering work of Cancho and Solé [6] who attempt to derive Zipf law using the principle of least action', while Hunt [10] refers to the work of Ferrer i Cancho and Solé in listing *minimum effort* among six underlying mechanisms that may lead to power laws—in addition to nonlinear dynamics (chaos), self-organized criticality, hierarchical dynamics, highly optimized tolerance, and fractal fracture properties. The contribution of the work of Ferrer i Cancho and Solé was acknowledged in other diverse contexts: quantitative linguistics [11], public transport [12], immune networks [13], genetic coding [14], etc.

The proposed model [6] is based on the optimality principle, namely the principle of least effort in communications. Specifically, as the goal of language is communication, the efficiency or accuracy of communications is the subject of optimization. A set of n signals S and a set of m objects R are used to describe signals between a 'speaker' (sender) and a 'hearer' (receiver), and the objects of reference. The relation between S and R is modelled using a binary matrix A, where an element $a_{i,j} = 1$ if and only if signal s_i refers to object r_j. The model allows one to represent both *polysemy* (that is, the capacity for a signal to have multiple meanings by referring to multiple objects), and *synonymy* where multiple signals refer to the same object (figure 1).

J. Stat. Mech. (2010) P11025

The effort for the sender is low if the signal entropy is low, implying a high amount of ambiguity. H_S expresses the effort of the sender, between 0 and 1, via the log with respect to n:

$$H_S \equiv H_n(S) = -\sum_{i=1}^{n} p(s_i) \log_n p(s_i). \tag{1}$$

Conversely, the effort for the receiver to decode a particular signal s_i is small if there is little ambiguity, i.e. the probability of a signal s_i referring to one object r_j is high. In [6], this is expressed by the conditional entropy

$$H_{R|s_i} \equiv H_m(R|s_i) = -\sum_{j=1}^{m} p(r_j|s_i) \log_m p(r_j|s_i). \tag{2}$$

The effort for the receiver is then dependent on the probability of each signal and the effort to decode it, that is

$$H_{R|S} \equiv H_m(R|S) = \sum_{i=1}^{n} p(s_i) H_{R|s_i}. \tag{3}$$

When this entropy is minimal, i.e. there is a one-to-one mapping between signals and objects, this effort is minimal. In computing the probabilities, we use the following:

$$p(s_i|r_j) = \frac{a_{i,j}}{\omega_j}, \tag{4}$$

where ω_j is the number of synonyms for object r_j, that is $\omega_j = \sum_i a_{i,j}$. That is, the probability of using a synonym is equally distributed over all synonyms referring to a particular object. Importantly, it is also assumed that $p(r_j) = 1/m$ is uniformly distributed over the objects, leading to a joint distribution:

$$p(s_i, r_j) = p(r_j)p(s_i|r_j) = \frac{a_{i,j}}{m\omega_j}. \tag{5}$$

A cost function Ω_λ is introduced to combine the effort of sender and receiver, with $0 \leq \lambda \leq 1$ trading off the effort between sender and receiver as follows:

$$\Omega_\lambda = \lambda H_{R|S} + (1 - \lambda) H_S. \tag{6}$$

In this representation, the binary matrix A is the variable of optimization, and we minimize the cost Ω_λ for different values of λ. In the extreme cases only the sender's effort ($\lambda = 0$) or the receiver's effort ($\lambda = 1$) is considered.

Comment 1. It should be noted that the cost function Ω_λ given by (6) is a specific case of a more general *energy* function that a communication system must minimize [7]:

$$\Omega^0_\lambda = -\lambda I(S; R) + (1 - \lambda) H_S, \tag{7}$$

where $I(S; R) = H_R - H_{R|S}$ is the mutual information. As pointed out by Ferrer i Cancho [7] in a comprehensive follow-up study, communicative efficiency is totally favoured when $\lambda = 1$, while saving cost is totally favoured when $\lambda = 0$. In addition, as mentioned in another follow-up study of Ferrer i Cancho and Díaz-Guilera [8], this energy function better accounts for subtle communication efforts, noting that $H(S)$ is both a

J. Stat. Mech. (2010) P11025

source of effort for the sender and the receiver because the word frequency affects not only word production but also recognition of spoken and written words. The component $I(S;R)$ also implicitly accounts for both $H_{S|R}$ (a measure of the sender's effort of coding objects) and $H_{R|S}$ (i.e., a measure of the receiver's effort of decoding signals). We follow the 'least-effort communication' terminology rather than describing the efforts via the energy consumed by the sender and the receiver, as the 'least effort' is a more accepted term in computational linguistics. One may also point out an interpretation of 'least effort' as the effort spent in order to transmit a bit of the information, i.e., consider a cost function in the form $\Omega_\lambda/I(S;R)$, but this possibility is out of scope of this study.

It follows that

$$\Omega^0_\lambda = -\lambda H_R + \lambda H_{R|S} + (1-\lambda)H_S = -\lambda H_R + \Omega_\lambda. \tag{8}$$

This more generic representation makes clear that the cost function Ω_λ is suitable when $H(R)$ is constant, and the uniformity condition $p(r_j) = 1/m$ ensures precisely that.

It is, of course, clear that the uniformity condition is a strong assumption that many developed natural languages do not satisfy across their full vocabularies, or even within nouns. For instance, Ferrer i Cancho [7] commented that 'the word dog is more likely to be used than the word aardvark because, roughly speaking, aardvarks, edentate mammals that are common in Southern Africa, have a much more restricted habitat than dogs'. The study of Ferrer i Cancho [7] replaced the uniformity assumption with $p(r_j) = \omega_j/M$, where, as defined above, ω_j is the number of synonyms for object r_j, while M is the total amount of synonyms, $M = \sum_{j=1}^{m} \omega_j$. This extension removed the constraint that $H(R)$ is constant, focusing on the energy function Ω^0_λ.

In our study we nevertheless analyse the more simple case where $H(R)$ is constant, attempting to analytically derive optimal solutions and the ensuing dynamics. In addition, the following section points out some important similarities between the models using $p(r_j) = (\omega_j/M)$ (model **A**) and $p(r_j) = (1/m)$ (model **B**).

Finally, we note that the *accuracy of the communication* as the mutual information $I(S;R)$ is used to measure the result of the trade-off between these efforts. Matrices were evolved to minimize cost Ω_λ by a simple mutation-based genetic algorithm (GA) [6]. Each computational experiment employed a 'greedy' strategy: whenever the cost of a candidate solution was smaller than the current minimal cost, the solution was accepted—otherwise, it was rejected. The algorithm was stopped when there was no progress during a given number of generations.

2. Recapitulation of the results

The information-theoretic model has been shown by Ferrer i Cancho and Solé to generate a phase transition in $I(S;R)$ at a critical value of $\lambda^* \approx 0.4$ where the efforts of the sender and receiver were argued to be balanced. It also produced frequency distributions for signals ranked (sorted) by decreasing frequencies that follow power laws (e.g., Zipf's law: $P(\rho) \propto \rho^{-\alpha}$ with $\alpha \approx 1$, ρ denoting the rank, and $P(\rho)$ denoting the frequency), for the matrices that corresponded to the critical λ^*.

We implemented and verified the original method of Ferrer i Cancho and Solé tracing the accuracy of the communication $I(S;R)$, as a function of λ, for 150×150 matrices. We

J. Stat. Mech. (2010) P11025

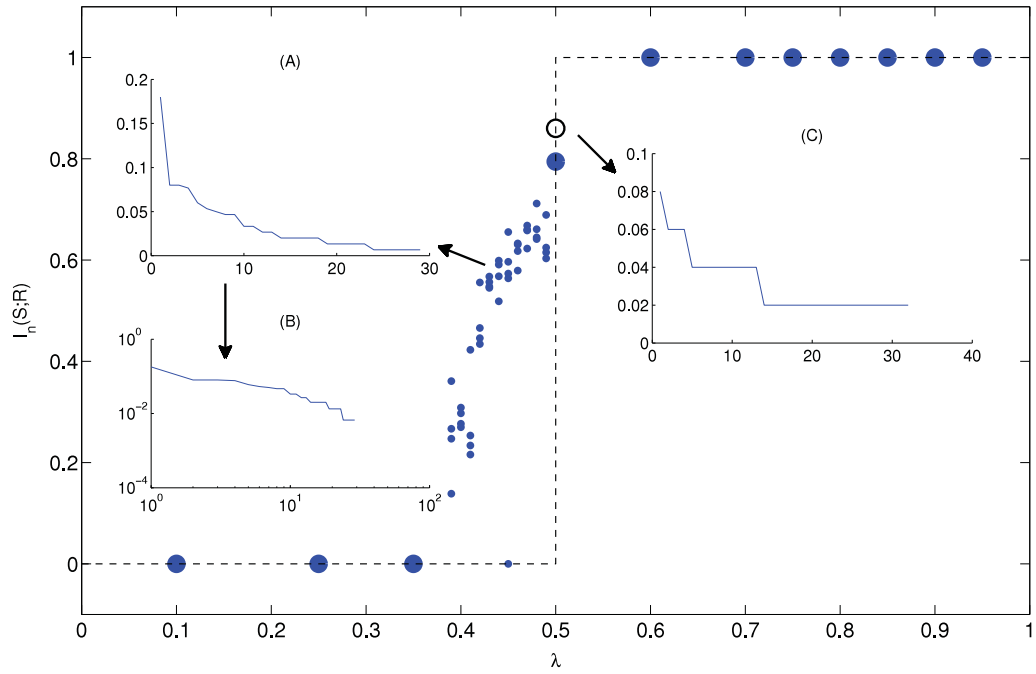

Figure 2. The mutual information as function of λ. Analytical results: dashed line as a step function. GA results: solid blue circles, with average values only for $\lambda < 0.38$ or $\lambda > 0.5$. A: power law observed for a GA solution, local minimizer, at $\lambda = 0.44$ and $I(S;R) \approx 0.59$, for a 150×150 matrix. B: the corresponding power law on a log–log scale. C: an inverse-factorial (sub-logarithmic) law for a global minimizer, marked ○, at $\lambda = 0.5$ and $I(S;R) \approx 0.86$, for a 50×50 matrix.

observed (figure 2) that for small values $\lambda < \lambda^*$, $I(S;R)$ is equal to (or near) zero, before undergoing a transition in the vicinity $\lambda \approx \lambda^*$. Single-signal systems dominate for $\lambda < \lambda^*$: 'because every object has at least one signal, one signal stands for all the objects' [6]. Low $I(S;R)$ indicates that the system is unable to convey information in this domain. Rich vocabularies are found after the transition, for $\lambda > \lambda^*$. Full vocabularies are attained for very high λ. The maximal value of $I(S;R)$ indicates that the associations between signals and objects are one-to-one maps, removing any redundancy in the vocabulary.

It has been recently pointed out by Ferrer i Cancho and Díaz-Guilera [8] that the optimal solutions found by this method are, however, only local minima. Ferrer i Cancho and Díaz-Guilera analytically derived global minima of the cost Ω_λ, showing that the phase transition is in fact a step function, completely separating two domains, $\lambda < \lambda^*$ and $\lambda > \lambda^*$, by the transition point $\lambda = \lambda^*$. Moreover, they proved for the family of solutions that satisfy the assumption $p(r_j) = 1/m$ (model **B**) that (i) the only global minimizers for the first domain ($\lambda < \lambda^*$) are given by single-signal communication systems—that is, one signal refers to all objects; (ii) the only global minimizers for the transition point ($\lambda = \lambda^*$) are given by matrices where no two signals refer to the same object; and (iii) the only global minimizers for the second domain ($\lambda > \lambda^*$) are given by matrices with one-to-one mapping between signals and objects (if $n = m$), while if $n > m$ every signal has to refer to at most one object and all objects are referred to, and if $n < m$ and the ratio m/n

J. Stat. Mech. (2010) P11025

is an integer, the signals refer to the same number m/n of objects and all objects are referred to.

Similar results were obtained for model **A**, i.e., $p(r_j) = \omega_j/M$. Specifically, (i) the only global minimizers for the first domain ($\lambda < \lambda^*$) are also given by single-signal communication systems—the difference from model **B** being that one signal refers to (*a subset of*) all objects; (ii) the only global minimizers for the transition point ($\lambda = \lambda^*$) are again given by matrices where no two signals refer to the same object; the only difference from model **B** is that some objects may have no signals at all; (iii) the only global minimizers for the second domain ($\lambda > \lambda^*$) are given by matrices with one-to-one mapping between signals and objects (if $n = m$), while for $n \neq m$ the signals refer to the same non-zero number of objects, and every object is referred to by at most one signal.

In short, for both models (**A** and **B**): the first domain ($\lambda < \lambda^*$) is characterized by single-signal communication systems; and the second domain ($\lambda > \lambda^*$) is characterized by one-to-one mapping between signals and objects (or, for $n \neq m$, solutions that maximally contain such one-to-one mapping). The transition point allows the solutions from either of these domains, as well as any solution without synonyms. Crucially, the global optima preclude solutions with synonyms for any λ.

The resulting similarity between models **A** and **B** in terms of the global minimizers further justifies the choice of a simpler model **B** for a detailed analysis motivated in section 3, while leaving analysis of model **A** to a future study.

3. Motivation

These observations demonstrated that the computational method employed by Ferrer i Cancho and Solé [6] does not reach the global optimum. Our computational experiments showed that the sharpness of the phase transition in the accuracy of the communication $I(S;R)$ as a function of λ is dependent on the overall computational effort. That is, the more iterations allowed within the algorithm, the sharper the transition appears. Specifically, the critical value $\lambda^* \approx 0.44$ was slightly higher than $\lambda^* \approx 0.4$ reported by Ferrer i Cancho and Solé [6]. The follow-up study of Ferrer i Cancho [7] identified the critical $\lambda^* \approx 0.5$. Both studies noted the emergence of power laws as a result of averaging multiple solutions. However, the individual global minimal solutions (i.e., individual matrices) at the phase transition do not necessarily exhibit power laws as their frequency distributions.

It is important to point out that any power law under consideration is the frequency distribution of an individual solution (a minimizer), and not a power-law divergence of some order parameter in the vicinity of the phase transition. In other words, the mechanism (i.e., the least-effort communications principle) is a candidate to generate power laws *within* minimizers that correspond to the phase transition, but not a candidate to explain the power-law divergence of an order parameter (e.g., characteristic length) at the critical value of some control parameter.

The absence of power-law frequency distributions within individual global minimizers strongly motivates a further study. Firstly, we observe that there are, in general, multiple values of the accuracy of the communication $I(S;R)$ for a given cost Ω_λ. That is, there are multiple *minimizers*, i.e. matrices A that obtain the same cost Ω_λ, but differ in the corresponding values of $I(S;R)$. Secondly, the fact that the local minimizers

J. Stat. Mech. (2010) P11025

found computationally do exhibit power laws, while the theoretical global minimizers do not, puts under question the mechanism behind the emergence of power laws in this model. Finally, one may wish to explore alternative forms that dominate the frequency distributions of global minimizers at $\lambda = \lambda^*$.

The phase transition is the focus of our investigation. The original study of Ferrer i Cancho and Solé did not interpret the accuracy of the communication $I(S;R)$ as some kind of a macroscopic (order) parameter. One may see, however, that when the trade-off parameter λ decreases, the inverted accuracy $1-I(S;R)$ undergoes the transition, as shown in figure 2. That is, the inverted accuracy $1-I(S;R)$ may be interpreted as an order parameter attaining the maximum value of 1 for the single-signal systems that can be seen as having maximal polysemy, being completely 'ordered', and the minimum value of 0 for the one-to-one maps that lack any polysemy, being completely 'disconnected'.

4. Results

4.1. Global minimizers

The reason that the computational experiment [6] does not find the global theoretical minima is the extreme complexity of the search-space, and the 'greedy' nature of the mutation-driven genetic algorithm. The solutions with synonyms may only be local minimizers, where the algorithm would be trapped by the low probability of simultaneous mutations required to jump to a better candidate. This is the explanation for the increasing sharpness of the transition that was obtained with a higher computational effort—that made it likelier to escape some local minima. The study of Ferrer i Cancho and Díaz-Guilera [8] obtained and characterized the global minima, but constrained the minimization space to a discrete space, i.e. binary matrices, rather than the continuous space of probability distributions. The notion of locality in the discrete space in terms of Hamming distance is not necessarily compatible with the notion of locality in the continuous space. In order to verify that solving in the continuous space does not change the global minima, we provide an analytical solution of the problem in the continuous space that generalizes the results of Ferrer i Cancho and Díaz-Guilera, for the model **B**.

Henceforth, we consider a joint probability distribution p containing joint probabilities $p(s_i, r_j)$ as the object of minimization, given the cost function Ω_λ, staying within the model **B**.

First of all, we establish the following result that applies to local minimizers (clearly including global minimizers).

Lemma 1. *Each solution (the joint probability distribution p) locally minimizing the function Ω_λ, specified by the equation (6), $0 \leq \lambda \leq 1$, can be represented as a function $f : R \to S$ such that*

$$p(s_i, r_j) = \begin{cases} 1/m & \text{if } s_i = f(r_j); \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The proof is given in appendix A. Note that each solution, i.e., each distribution p, corresponds via expression (5) to a matrix A (henceforth called the *minimizer matrix*) which is given in terms of the function f as follows:

$$a_{i,j} = \begin{cases} 1 & \text{if } s_i = f(r_j); \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The main outcome of this observation is that the analytical minimization of the suggested cost function results in solutions without synonyms—since any function f precludes multiple signals s referring to the same object r. That is, each column in the minimizer matrix has precisely one non-zero element. Polysemy is allowed within the solutions. Importantly, the representation of the solutions as functions subsumes the classes of global solutions described by Ferrer i Cancho and Díaz-Guilera [8]. The local minimizers obtained by GA are obviously not the local minimizers in continuous space, but we focus on global minimizers henceforth. The global minimizers in both discrete and continuous space are the same.

Secondly, we make the following observation.

Lemma 2. *For each solution p minimizing the function Ω_λ,*

$$H_{R|S} + \frac{1}{\log_n m} H_S = 1. \tag{11}$$

The proof is given in appendix B.

Corollary 3. *If $n = m$, $H_{R|S} + H_S = 1$.*

It follows that for $n = m$ the joint entropy $H_{S,R} = H_{R|S} + H_S = 1$. Although the best trade-off for the global solutions depends on λ, the actual effort values may be quite different. The last lemma and corollary inform that the actual values H_S and $H_{R|S}$ are interrelated (e.g., if the sender's effort is high, the receiver's effort is low), while the joint entropy (as a proxy of the joint effort) is kept fixed, more precisely at its mid-point for square matrices, or at some point skewed by $1/\log_n m$ otherwise.

Corollary 4. *If the entropies $H_{R|S}$ and H_S are equal, then*

$$H_{R|S} = H_S = \frac{\log_n m}{\log_n m + 1}.$$

Thirdly, using these results, we obtain analytical solutions for the minimization of the function Ω_λ, that depend on the critical value of the parameter λ:

$$\lambda^* = \frac{\log_n m}{\log_n m + 1}. \tag{12}$$

Lemma 5. *If $\lambda < \lambda^*$, then p minimizing the function Ω_λ is given by*

$$p(s_i, r_j) = \begin{cases} \dfrac{1}{m} & \text{for a particular } i^* \text{ and all } j; \\ 0 & i \neq i^* \text{ and all } j. \end{cases}$$

If $\lambda > \lambda^*$, then p minimizing the function Ω_λ is given by

$$p(s_i, r_j) = \begin{cases} \dfrac{1}{m} & \text{for } i^* \text{ and } j^* \text{ where } s_i^* = f(r_j^*); \\ 0 & \text{otherwise} \end{cases}$$

for some function $f : R \to S$, subject to

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) = \frac{1}{n}.$$

If $\lambda = \lambda^*$, then any p representable as a function $f : R \to S$ as specified by (9) minimizes the function Ω_λ.

The proof is given in appendix C.

Corollary 6. *If $n = m$, $\lambda^* = \frac{1}{2}$.*

The results of Ferrer i Cancho and Díaz-Guilera [8] may now be derived in the discrete space of binary matrices. In particular, for $n = m$:

- for all values $\lambda < \lambda^*$, the only minimizers A allowed are the single-signal matrices with a single row i^* of $a_{i^*,j} = 1$ for any j, and any of these produces $I(S; R) = 0$;
- for all values $\lambda > \lambda^*$, the only minimizers A allowed are the matrices with one-to-one mapping, i.e., a single $a_{i^*,j^*} = 1$ in each row, and any of these produces $I(S; R) = 1$;
- for the critical value λ^*, there are multiple different minimizers A such that no two signals refer to the same object (but one signal may refer to multiple objects), that cover the range $0 \leq I(S; R) \leq 1$.

The results for $n \neq m$ [8] are also easily derived, but are omitted due to the lack of space. This confirms that the theoretical phase transition is indeed a step function (figure 2).

This still does not answer the main question on the emergence of power laws in our model system. Obviously, the fact that human languages do manifest power-law distributions is not debated here—we simply point out that the interpretation of the least-effort communication principle, modelled in the considered way, is inadequate when one seeks global optima, as was indicated by Ferrer i Cancho and Díaz-Guilera [8]. The fact that the local optima found by the simulation studies do exhibit power laws in the frequency distributions may indicate that the evolution of languages in Nature is likely to be trapped in local optima for long periods, and specifically that 'the need for communicating (the need for $I(S, R) > 0$) may be a serious obstacle for human language reaching the global optimum' [8]. Nevertheless, another intriguing possibility is that alternative forms dominate frequency distributions of the global minimizers.

4.2. Power laws

The analysis presented above does not single out any of the global minimizers for the critical value λ^*: all of these are equal in attaining the minimum cost Ω_{λ^*}. The scale-free solutions may play some special role among the minimizers that 'co-exist' at $\lambda = \lambda^*$ if the optimization task is modified. This subsection explores one such possibility.

J. Stat. Mech. (2010) P11025

Lemma 7. *The following condition is necessary for $P(\rho) \propto 1/\rho$, where ρ is the rank of the frequency distribution, for $n \to \infty$, assuming that $\log_n m$ is finite:*

$$\frac{2\log_n m - 1}{\log_n m} H_S = H_{R|S}. \tag{13}$$

The proof is given in appendix D. In particular, it derives the equation

$$H_S = -\frac{A_n \ln(1/A_n) + \gamma_1(n+1) - \gamma_1}{A_n \ln n}, \tag{14}$$

where A_n is the nth harmonic number, $\gamma_1 = -0.072\,815\,845\,48\cdots$ is a Stieltjes constant, and $\gamma_1(n+1)$ is the generalized Stieltjes constant (see appendix D for more details). Specifically, as n grows, the entropy H_S approaches $\frac{1}{2}$ from above. When m is of the order of $\sqrt{n}$, i.e. $\log_n m = \frac{1}{2}$, the entropy $H_{R|S}$ asymptotically vanishes.

Corollary 8. *If $n = m$, the following condition is necessary for $P(\rho) \propto 1/\rho$, where ρ is the rank of the frequency distribution, for $n \to \infty$:*

$$H_S = H_{R|S}. \tag{15}$$

This establishes that a power-law frequency distribution asymptotically leads to a precise balance between the two involved efforts. We have not established that this balance is a sufficient condition for the emergence of power laws. Nevertheless, the condition points out that when a frequency distribution of a minimizer obeys a power law, then in addition to minimizing the cost, the efforts of the sender and receiver are equal for large systems with $n \to \infty$ and $m \to \infty$.

It should be pointed out for real-world human language communications (e.g. $n \approx m \approx 10^6$; for instance, the number of English words in the Oxford English Dictionary is about 600 000), the balance is not achieved by an equal split of the effort, but is rather given by $H_S \approx 0.67$ and $H_{R|S} = 0.33$, obtained using (14) derived in appendix D, and lemma 2. In other words, a power-law frequency distribution that minimizes the combined effort Ω_λ requires that the sender spends about twice as much effort as the receiver. It may be argued that 10^6 words is not sufficiently large a vocabulary for the equal split $H_S = H_{R|S}$.

4.3. Configurations

In the remainder of the paper, we study alternative forms that dominate frequency distributions of the global minimizers, and carry out some more detailed analysis of the corresponding minimizers.

Definition 9. *The configuration for an $(n \times m)$ minimizer matrix A is an $(m+1)$-dimensional vector $\pi = (\pi_0 \cdots \pi_k \cdots \pi_m)$, where $0 \le k \le m$ and π_k are non-negative integers, such that there are π_k rows with k non-zero elements $a_{i,j} \neq 0$ in the matrix.*

For example, if there are five rows with a single non-zero element, that is, there are five signals each of which refers to only one object, then $\pi_1 = 5$ (one may say that there are five 'singles' in the matrix). Informally, a configuration is a histogram of signal usage, ordered by the number of referred objects, i.e. ranging from 0 to m—hence, $(m+1)$ dimensions in a configuration vector. A similar representation is discussed by Trosso [15].

Multiple minimizer matrices may share the same configuration. In fact, a configuration defines an equivalence class for minimizer matrices. It is clear that the configuration for an $(n \times m)$ minimizer matrix A satisfies the constraints $\sum_k \pi_k = n$, and $\sum_k k\pi_k = m$. The first condition, $\sum_k \pi_k = n$, ensures that the number of rows in the matrix described by the configuration is n, and the second condition, $\sum_k k\pi_k = m$, ensures that the number of non-zero elements is precisely m. We shall refer to minimizers in the same equivalence class as configuration instances.

Example 10. Consider (3×3) minimizer matrices A. The first configuration $(2; 0; 0; 1)$ describes all matrices with two rows containing only zero elements, and a single row, for some i^*, of three elements $a_{i^*,j} = 1$ for any j. There are three such matrices obtained by permuting the rows. The second configuration $(0; 3; 0; 0)$ describes all six matrices with three rows, each containing a single $a_{i^*,j^*} = 1$. The third configuration $(1; 1; 1; 0)$ describes all 18 matrices with one row containing only zero elements, a single row with one element $a_{i,j} = 1$, and a single row with two elements $a_{i,j} = 1$.

We shall denote d consecutive zeros in a configuration by $[0]^d$, so $(0; 3; 0; 0)$ is equivalent to $(0; 3; [0]^2)$.

A configuration vector π defines a mapping from minimizers to their configuration, i.e. $\pi : S^R \to [0, n]^{m+1}$. Here S^R is the set of all functions $f : R \to S$ that characterize minimizers. Each specific configuration vector π maps a subset of functions from S^R to the single corresponding configuration in $[0, n]^{m+1}$. For example, the functions $f : R \to S$ that map all objects to the same signal (producing the minimizer matrices with a single row i^* of three elements $a_{i^*,j} = 1$) are mapped to the first configuration $(2; 0; 0; 1)$ in the example above.

For a fixed λ, all minimizers that are mapped to the same configuration (i.e., the configuration instances) obtain the same cost Ω_λ. The configuration instances also always agree on accuracy $I(S; R)$.

Lemma 11. *Any $(m + 1)$-dimensional vector $\pi = (\pi_0 \cdots \pi_k \cdots \pi_m)$, where π_k are non-negative integers, that satisfies the constraints $\sum_k \pi_k = n$, and $\sum_k k\pi_k = m$ is a configuration.*

This observation establishes that for any vector satisfying these constraints there necessarily exists a minimizer matrix with π_k rows with k non-zero elements $a_{i,j} \neq 0$, where $0 \leq k \leq m$. In general, there are multiple such matrices.

While there are multiple different minimizers A for the critical value λ^* that cover the range $0 \leq I(S; R) \leq 1$, their distribution across the configurations is not uniform. Some of the configurations have more instances (cf example 10).

Every matrix has a frequency distribution, and all the instances of a configuration share a frequency distribution.

Example 12. The configuration $(2; 3; 2; [0]^5)$ describing an equivalence class for some (7×7) minimizer matrices has more instances than any other. The frequency distribution shared by all the instances of this configuration contains $p(s_1) = p(s_2) = \frac{2}{7}$ (for the two rows with two non-zero elements: $\pi_2 = 2$); $p(s_3) = p(s_4) = p(s_5) = \frac{1}{7}$ (for the three rows with one non-zero element: $\pi_1 = 3$); and $p(s_6) = p(s_7) = 0$ (for two rows with all zero elements: $\pi_0 = 2$).

J. Stat. Mech. (2010) P11025

4.4. Maximizing the number of instances

We shall now derive an analytical representation for the configuration that describes minimizers at the phase transition and has a maximal number of instances (i.e., for the most populous configuration).

Lemma 13. *The number of matrices $\mathcal{L}$ described by the configuration vector $\pi = (\pi_0 \cdots \pi_k \cdots \pi_m)$, where π_k are non-negative integers, $0 \leq k \leq m$, is*

$$\mathcal{L}(\pi) = \frac{n!m!}{\prod_{k=0}^{m} \pi_k!(k!)^{\pi_k}}. \tag{16}$$

The proof and example are given in appendix E. Let us briefly explain the terms of the expression (16). The overall number of different instances for a configuration is simply

$$\mathcal{L}(\pi) = \mathcal{L}_s(\pi) \cdot \mathcal{L}_r(\pi),$$

where

$$\mathcal{L}_s(\pi) = \frac{n!}{\prod_{k=0}^{m} \pi_k!} \tag{17}$$

is the number of permutations of matrix rows, and

$$\mathcal{L}_r(\pi) = \frac{m!}{\prod_{k=0}^{m} (k!)^{\pi_k}} \tag{18}$$

captures the number of possibilities to permute 'ones' across an individual row, that is, to permute across m columns.

The expression (16) allows us to approach the question of finding the vector π^* that maximizes $\mathcal{L}(\pi)$ under the conditions $\sum_{k=0}^{m} \pi_k = n$ and $\sum_{k=1}^{m} k\pi_k = m$.

Lemma 14. *The vector $\pi^* = (\pi_0^* \cdots \pi_k^* \cdots \pi_m^*)$, where π_k^* are real numbers, that maximizes $\mathcal{L}(\pi)$ under the conditions $\sum_{k=0}^{m} \pi_k = n$ and $\sum_{k=1}^{m} k\pi_k = m$, asymptotically follows the Poisson distribution with the average (m/n) being multiplied by n:*

$$\pi_k^* \approx \frac{n\mathrm{e}^{-(m/n)}(m/n)^k}{k!}. \tag{19}$$

The proof is given in appendix F. For example, if $m = n$, the optimal vector is

$$\pi_k^* = \frac{n}{ek!}. \tag{20}$$

Let us exemplify this solution for an (50×50) minimizer, where the configuration $(18; 19; 9; 3; 1; [0]^{46})$ was found to be the most populous by an explicit calculation of expression (16) for all configurations. The solution (20) suggests the vector $\pi^* \approx (18.3940; 18.3940; 9.1970; 3.0657; 0.7664; \ldots)$, which differs only slightly due to the non-integer nature of the solution. The integer variations around the vector π^* make the dominance of this configuration even more significant. A contrasting example of a configuration that corresponds to a power-law frequency distribution is provided at the end of section 4.5.

J. Stat. Mech. (2010) P11025

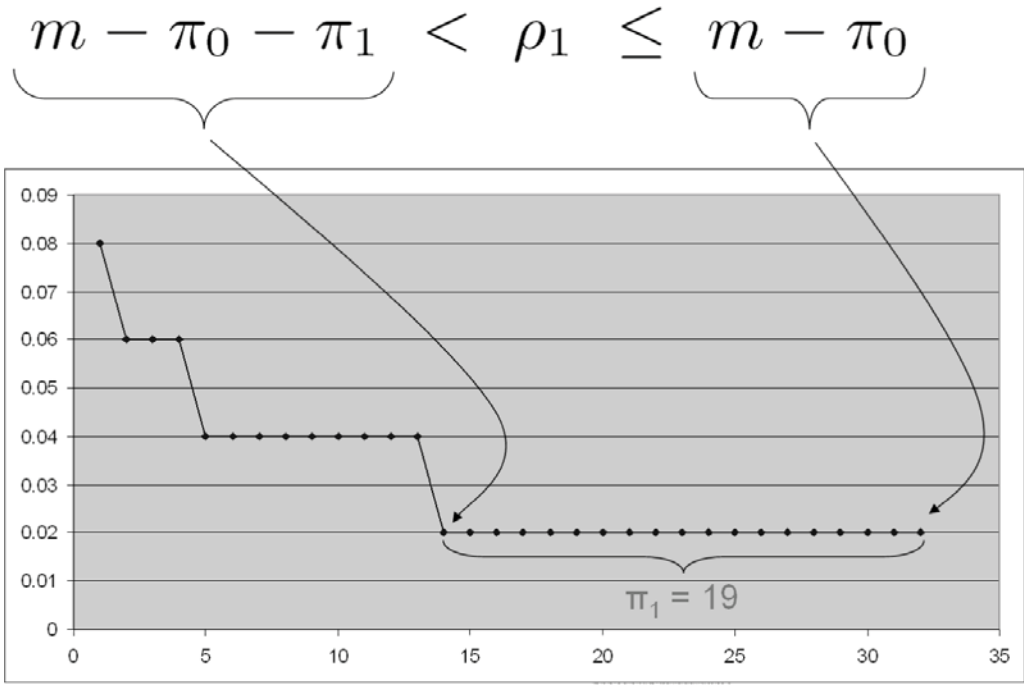

Figure 3. The frequency distribution (vertical axis) and the rank (horizontal axis).

4.5. Inverse-factorial law

We have determined that the most populous configurations are integer variations around the vector π_k^* that follows the Poisson distribution with the average (m/n) being multiplied by n, e.g. given by (20) if $m = n$. Henceforth we consider square matrices, $n = m$.

Let us consider the rank ρ_k of the sequence of signals s_i sorted by their frequency $P(\rho_k) \equiv k/m$, where $\rho = 1$ denotes the highest rank (highest frequency). The rank ρ_k satisfies the following condition (cf figure 3):

$$m - \sum_{j=0}^{k} \pi_j < \rho_k \leq m - \sum_{j=0}^{k-1} \pi_j. \tag{21}$$

Expressing k as a function of the rank ρ_k, and setting $P(\rho_k) = k/m$, yields (appendix G) the inverse-factorial dependency

$$P(\rho_k) \approx \frac{1}{m} \Gamma^{-1}\left(\frac{m e^{\xi-1}}{\rho_k}\right), \tag{22}$$

where $\Gamma^{-1}(x)$ is the principal branch of the inverse $\Gamma(x)$ function, $\Gamma(x) = (x+1)!$ [16] (cf appendix G), and $0 < \xi \leq \ln(e-1)$. This dependency reduces the symbols' frequency much slower than a power law $P(\rho_k) \propto \rho_k^{-\alpha}$—in fact, the rate of change is sub-logarithmic.

The frequency distribution of the configuration $(18; 19; 9; 3; 1; [0]^{46})$, minimizing in the space of (50×50) matrices, obeys such a sub-logarithmic law (i.e., inverse-factorial law) with the highest rank $\rho_4 = 1$. That is, the sole signal referring to four objects $(\pi_4 = 1)$ is the most frequent.

This establishes that at the phase transition the space of minimizers is dominated by inverse-factorial (sub-logarithmic) rather than power laws. The latter type is not ruled out completely. For instance, the configuration $(35; 5; 5; 1; 1; 1; [0]^2; 1; [0]^6; 1; [0]^{35})$, that also describes minimizers in the space of (50×50) matrices, has a frequency distribution closely following a power-law distribution with the highest rank $\rho_{15} = 1$. That is, the most frequent signal refers to 15 objects ($\pi_{15} = 1$). However, this configuration has $\approx 10^{21}$ times fewer instances than the configuration $(18; 19; 9; 3; 1; [0]^{46})$.

5. Conclusions

In this paper we critically examined an information-theoretic model proposed by Ferrer i Cancho and Solé [6] in the attempt to formalize the principle of least effort in communications. The model suggests minimizing the overall cost Ω_λ balanced between the speaker effort and listener effort, with some trade-off λ. When the task is solved computationally by a 'greedy' search method such as a mutation-based genetic algorithm, the model appears sufficiently rich to explain the emergence of power laws (specifically, Zipf's law) in human languages. The solutions minimizing the balanced cost Ω_λ are characterized by frequency distributions of the signals that refer to (possibly multiple) objects in various ways. Specifically, Ferrer i Cancho and Solé [6] observed one prominent frequency distribution—Zipf's law—in the vicinity of the transition between referentially useless systems (one signal for all referable objects) and indexical reference systems (one signal per object). The phase transition, during which Zipf's law is found, is defined in the space of communication accuracy (information content I) expressed in terms of the trade-off parameter λ.

We also followed up on a recent study of Ferrer i Cancho and Díaz-Guilera [8] who proved that the optimal solutions found by the computational method [6] are only local minimizers, not reaching the global minima. The analytically derived global minima of the cost Ω_λ produce the phase transition as a step function [8]. Most importantly, the global minimizers at the phase transition do not necessarily exhibit power laws.

Our investigation focused on the phase transition between referentially useless systems and indexical reference systems, trying to clarify the mechanism behind the emergence of power laws in the original model, as well as to explore alternative forms of frequency distributions that occur within the global minimizers at $\lambda = \lambda^*$.

In doing so, we explicitly solve the *continuous* optimization problem, and subsume the more specific result of Ferrer i Cancho and Díaz-Guilera [8] obtained within a discrete space. The new results contrast Zipf's law found computationally (for local minima) in the *vicinity* of the phase transition, with an inverse-factorial (sub-logarithmic) law found *at* the transition that corresponds to global minima. The inverse-factorial law is observed to be the most dominant, i.e. occurring in solutions that were significantly more widespread at the transition.

We reiterate that we do not debate here that human languages manifest power-law distributions, but point out that the information-theoretic interpretation of the least-effort communication principle [6] is not sufficiently strong for generating power laws at the global minima of the effort (unless some additional constraints are imposed—cf section 4.2). The study of Ferrer i Cancho [7] which replaced the uniformity assumption and proposed a more generic energy (cost) function deserves a more detailed analysis

J. Stat. Mech. (2010) P11025

with respect to forms of frequency distributions that dominate the transition point. Nevertheless, we hope that the results reported here will not only reinvigorate the search for a more precise model capturing the principle of least effort and power laws, but also help to uncover and interrelate diverse critical phenomena that exhibit (sub-)logarithmic laws.

Acknowledgments

Mikhail Prokopenko is grateful for a 2009 Research Grant from The Max Planck Institute for Mathematics in the Sciences (Leipzig, Germany) on Information-driven Self-Organisation and Complexity Measures, and travel support from The University of Hertfordshire (UK). The authors thank CSIRO's Advanced Scientific Computing group for access to high performance computing resources used for simulation. Nihat Ay acknowledges travel support from CSIRO ICT Centre.

Appendix A

Lemma. *Each solution (the joint probability distribution P) locally minimizing the function Ω_λ, specified by the equation (6), $0 \leq \lambda \leq 1$, can be represented as a function $f : R \to S$ such that*

$$p(s_i, r_j) = \begin{cases} 1/m & \text{if } s_i = f(r_j); \\ 0 & \text{otherwise.} \end{cases} \tag{A.1}$$

In order to prove this lemma, we establish a few preliminary propositions[5].

A.1. Concavity

Consider a set $S = \{s_1, \ldots, s_n\}$ of signals with n elements and a set $R = \{r_1, \ldots, r_m\}$ of m objects, and denote with $\mathcal{P}(S \times R)$ the set of all probability vectors $p(s_i, r_j)$, $1 \leq i \leq n$, $1 \leq j \leq m$. We define the following functions on $\mathcal{P}(S \times R)$:

$$H_S(p) := -\sum_i p(s_i) \log_n p(s_i)$$

and

$$H_{R|S}(p) := -\sum_i p(s_i) \sum_j p(r_j|s_i) \log_m p(r_j|s_i).$$

Proposition 1. *The functions H_S and $H_{R|S}$ are concave in p.*

Proof. The statements follow from well-known convexity properties of the entropy and the relative entropy.

[5] These results were obtained by Nihat Ay.

(i) *Concavity of* H_S: we rewrite H_S as

$$H_S(p) = -\sum_i \left(\sum_j p(s_i, r_j) \right) \log_n \left(\sum_j p(s_i, r_j) \right).$$

The concavity of H_S follows directly from the concavity of the Shannon entropy.

(ii) *Concavity of* $H_{R|S}$: we rewrite the function $H_{R|S}$ as

$$\begin{aligned} H_{R|S}(p) &= -\sum_i p(s_i) \sum_j p(r_j|s_i) \log_m p(r_j|s_i) \\ &= -\sum_{i,j} p(s_i, r_j) \log_m \frac{p(s_i, r_j)}{\sum_j p(s_i, r_j)} \\ &= -\sum_{i,j} p(s_i, r_j) \log_m \frac{p(s_i, r_j)}{m\,(1/m) \sum_j p(s_i, r_j)} \\ &= -\sum_{i,j} p(s_i, r_j) \log_m \frac{p(s_i, r_j)}{(1/m) \sum_j p(s_i, r_j)} + 1. \end{aligned}$$

The concavity of $H_{R|S}$ now follows from the joint convexity of the relative entropy $(p, q) \mapsto D(p\|q) = \sum_{i,j} p(s_i, r_j) \log_m (p(s_i, r_j)/q(s_i, r_j))$.

□

With a number $0 \le \lambda \le 1$, we now consider the corresponding convex combination of the functions H_S and $H_{R|S}$:

$$\Omega_\lambda(p) = \lambda\, H_{R|S}(p) + (1 - \lambda)\, H_S(p).$$

From the above proposition, it immediately follows that Ω_λ is also concave.

Corollary 2. *The function* Ω_λ *is concave in* p.

A.2. Extreme points

We will consider the restriction of Ω_λ to the convex set

$$\mathcal{C} := \left\{ p \in \mathcal{P}(S \times R) : p(r_j) = \sum_i p(s_i, r_j) = \frac{1}{m} \text{ for all } j \right\}.$$

The extreme points of $\mathcal{C}$ are specified by the following proposition.

Proposition 3. *The set* $\mathcal{C}$ *has the extreme points*

$$\mathrm{Ext}(\mathcal{C}) = \left\{ p \in \mathcal{P}(S \times R) : p(s_i, r_j) = \frac{1}{m}\, \delta_{f(r_j)}(s_i) \right\},$$

where f *is a function* $R \to S$.

Proof. Consider the convex set

$$\mathcal{T} = \left\{ A = (a_{i|j})_{i,j} \in \mathbb{R}^{m \cdot n} : a_{i|j} \ge 0 \text{ for all } i, j, \text{ and } \sum_i a_{i|j} = 1 \text{ for all } j \right\}$$

of transition matrices. The extreme points of $\mathcal{T}$ are given by functions $f : j \mapsto i$. More precisely, each extreme point has the structure

$$a_{i|j} = \delta_{f(j)}(i).$$

Now consider the map $\varphi : \mathcal{T} \to \mathcal{C}$ that maps each matrix $A = (a_{i|j})_{i,j}$ to the probability vector

$$p(s_i, r_j) := \frac{1}{m}\, a_{i|j}, \qquad \text{for all } i, j.$$

This map is bijective and satisfies $\varphi((1-t)\,A + t\,B) = (1-t)\,\varphi(A) + t\,\varphi(B)$. Therefore, the extreme points of $\mathcal{C}$ can be identified with the extreme points of $\mathcal{T}$. □

A.3. Minimizers

If we now minimize the function Ω_λ over the set $\mathcal{C}$ then, with a local minimizer p, each further point q in the face $F(p)$ of $\mathcal{C}$ that contains p is a local minimizer with the same value. The following proposition shows that for $0 < \lambda < 1$, $F(p) = \{p\}$, which means that a minimizer is always an extreme point.

Lemma. *Let F and G be concave functions on a convex set $\mathcal{C}$, let $p_k \in \mathcal{C}$, $\alpha_k \in [0,1]$, $k = 1, \ldots, r$, satisfying $\sum_k^r \alpha_k = 1$. Then the equation*

$$(F+G)\left(\sum_k \alpha_k\, p_k\right) = \sum_k \alpha_k\, (F+G)(p_k) \tag{A.2}$$

implies

$$F\left(\sum_k \alpha_k\, p_k\right) = \sum_k \alpha_k\, F(p_k) \qquad \text{and} \qquad G\left(\sum_k \alpha_k\, p_k\right) = \sum_k \alpha_k\, G(p_k).$$

Proof. Without loss of generality, assume

$$F\left(\sum_k \alpha_k\, p_k\right) > \sum_k \alpha_k\, F(p_k).$$

Then

$$\begin{aligned}
(F+G)\left(\sum_k \alpha_k\, p_k\right) &= F\left(\sum_k \alpha_k\, p_k\right) + G\left(\sum_k \alpha_k\, p_k\right) \\
&> \sum_k \alpha_k\, F(p_k) + G\left(\sum_k \alpha_k\, p_k\right) \\
&\geq \sum_k \alpha_k\, F(p_k) + \sum_k \alpha_k\, G(p_k) \\
&= \sum_k \alpha_k\, (F+G)(p_k).
\end{aligned}$$

This is a contradiction to the equality (A.2). □

Proposition 4. *Let $0 < \lambda < 1$ and let p be a local minimizer of the map*

$$\mathcal{C} \to \mathbb{R}, \qquad p \mapsto \Omega_\lambda(p).$$

Then p is an extreme point of $\mathcal{C}$.

Proof. Consider a representation of p as a convex combination of points $p_k \in \mathrm{Ext}(\mathcal{C})$:

$$p = \sum_k \alpha_k\, p_k, \qquad \alpha_k > 0, \qquad \sum_k \alpha_k = 1.$$

We have to prove that $p = p_k$ for all k. This is done in several steps.

(1) The assumption that p is a local minimizer of Ω_λ implies

$$\begin{aligned}\Omega_\lambda(p) &= ((1-\lambda)\, H_S + \lambda\, H_{R|S})(p) \\ &= \sum_k \alpha_k\, ((1-\lambda)\, H_S + \lambda\, H_{R|S})(p_k) = \sum_k \alpha_k\, \Omega_\lambda(p_k).\end{aligned}$$

From the above lemma in appendix A.3 it therefore follows that

$$H_S(p) = \sum_k \alpha_k\, H_S(p_k) \qquad \text{and} \tag{A.3}$$

$$H_{R|S}(p) = \sum_k \alpha_k\, H_{R|S}(p_k). \tag{A.4}$$

(2) From the strict concavity of the entropy H_S with respect to the S-marginal we get

$$p(s_i) = p_k(s_i) \qquad \text{for all } k \text{ and } i. \tag{A.5}$$

If $p(s_i) > 0$ then (A.5) implies for all j

$$\begin{aligned}p(r_j|s_i) &= \sum_k \alpha_k\, \frac{p_k(s_i, r_j)}{p(s_i)} = \sum_k \alpha_k\, \frac{p_k(s_i, r_j)}{p_k(s_i)} \\ &= \sum_k \alpha_k\, p_k(r_j|s_i).\end{aligned} \tag{A.6}$$

(3) The function

$$\mathcal{T} \to \mathbb{R}, A = (a_{j|i})_{i,j} \mapsto - \sum_{\substack{i \\ p(s_i)>0}} p(s_i) \sum_j a_{j|i} \log_m a_{j|i}$$

is strictly concave. Together with (A.4) and (A.6) this implies that for all i with $p(s_i) > 0$

$$p(r_j|s_i) = p_k(r_j|s_i) \qquad \text{for all } j, k.$$

A combination with (A.5) yields

$$p(s_i, r_j) = p(s_i)\, p(r_j|s_i) = p_k(s_i)\, p_k(r_j|s_i) = p_k(s_i, r_j).$$

We finally observe that, also in the case $p(s_i) = 0$, the equality $p(s_i, r_j) = p_k(s_i, r_j)$ holds for all k and j:

$$0 \le p(s_i, r_j) \le \sum_{j'} p(s_i, r_{j'}) = p(s_i) = 0 \qquad \text{and also}$$

$$0 \le p_k(s_i, r_j) \le \sum_{j'} p_k(s_i, r_{j'}) = p_k(s_i) \overset{(A.5)}{=} p(s_i) = 0.$$

□

Consider the set of 0/1-matrices that have at least one '1'-entry in each column:

$$\mathcal{S} := \left\{ (a_{i,j}) \in \{0,1\}^{n\cdot m} : \sum_i a_{i,j} \geq 1 \text{ for all } j \right\}.$$

This set can naturally be embedded into the set $\mathcal{T}$, which we have considered in the proof of proposition 3:

$$\imath : \mathcal{S} \hookrightarrow \mathcal{T}, \qquad (a_{i,j})_{i,j} \mapsto a_{i|j} := \frac{a_{i,j}}{\sum_i a_{i,j}}.$$

Together with the map $\varphi : \mathcal{T} \to \mathcal{C}$ we have the injective composition $\varphi \circ \imath$. From proposition 3 it follows that the extreme points of $\mathcal{C}$ are in the image of $\varphi \circ \imath$. Furthermore, proposition 4 implies that all local, and therefore also all global, minimizers of Ω_λ are in the image of $\varphi \circ \imath$. The previous work of Ferrer i Cancho and Sole [6] refers to the minimization of the function

$$\tilde{\Omega}_\lambda := \Omega_\lambda \circ \varphi \circ \imath : \mathcal{S} \to \mathbb{R}.$$

It is not obvious how to relate local minimizers of this function, with an appropriate notion of locality in $\mathcal{S}$, to local minimizers of Ω_λ. However, we have the following obvious relation between global minimizers.

Corollary 5. *A point $p \in \mathcal{C}$ is a global minimizer of Ω_λ if and only if it is in the image of $\varphi \circ \imath$ and $(\varphi \circ \imath)^{-1}(p)$ globally minimizes $\tilde{\Omega}_\lambda$.*

Appendix B

Lemma. *For each solution p minimizing the function Ω_λ,*

$$H_{R|S} + \frac{1}{\log_n m} H_S = 1. \tag{B.1}$$

Proof. We begin by analysing the expression

$$H_{R|s_i} = -\sum_{j=1}^{m} p(r_j|s_i) \log_m p(r_j|s_i). \tag{B.2}$$

Using Bayes' rule, $p(r_j|s_i) = p(s_i, r_j)/p(s_i)$, we obtain

$$H_{R|s_i} = -\sum_{j=1}^{m} \frac{p(s_i, r_j)}{p(s_i)} \log_m \frac{p(s_i, r_j)}{p(s_i)}. \tag{B.3}$$

Expression (9) used within the logarithm yields

$$H_{R|s_i} = -\sum_{j=1}^{m} \frac{p(s_i, r_j)}{p(s_i)} \log_m \frac{1}{m p(s_i)}, \tag{B.4}$$

J. Stat. Mech. (2010) P11025

where the sum is taken for non-zero $p(s_i, r_j)$, while the sum's terms with $p(s_i, r_j) = 0$ are all equal to zero. It follows that

$$H_{R|s_i} = -\frac{1}{p(s_i)} \log_m \frac{1}{mp(s_i)} \sum_{j=1}^{m} p(s_i, r_j) = -\log_m \frac{1}{mp(s_i)}, \tag{B.5}$$

where the last reduction is obtained by using the marginalization $p(s_i) = \sum_j p(s_i, r_j)$. Hence,

$$H_{R|S} = \sum_{i=1}^{n} p(s_i) H_{R|s_i} = -\sum_{i=1}^{n} p(s_i) \log_m \frac{1}{mp(s_i)} \tag{B.6}$$

$$= \sum_{i=1}^{n} p(s_i)(\log_m m + \log_m p(s_i)) = 1 + \sum_{i=1}^{n} p(s_i) \log_m p(s_i) \tag{B.7}$$

$$= 1 + \frac{1}{\log_n m} \sum_{i=1}^{n} p(s_i) \log_n p(s_i) = 1 - \frac{1}{\log_n m} H_S.$$

The lemma's objective follows immediately. □

Appendix C

In this section, we establish the following lemma for the critical value

$$\lambda^* = \frac{\log_n m}{\log_n m + 1}. \tag{C.1}$$

Lemma. *If $\lambda < \lambda^*$, then p minimizing the function Ω_λ is given by*

$$p(s_i, r_j) = \begin{cases} \dfrac{1}{m} & \text{for a particular } i^* \text{ and all } j; \\ 0 & i \neq i^* \text{ and all } j. \end{cases}$$

If $\lambda > \lambda^*$, then p minimizing the function Ω_λ is given by

$$p(s_i, r_j) = \begin{cases} \dfrac{1}{m} & \text{for } i^* \text{ and } j^* \text{ where } s_i^* = f(r_j^*); \\ 0 & \text{otherwise} \end{cases}$$

for some function $f : R \to S$, subject to

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) = \frac{1}{n}.$$

If $\lambda = \lambda^*$, then any p, representable as a function $f : R \to S$ as specified by (9), minimizes the function Ω_λ.

Proof. Using the observation

$$H_{R|S} + \frac{1}{\log_n m} H_S = 1,$$

we express the receiver's effort as

$$H_{R|S} = 1 - \frac{1}{\log_n m} H_S$$

and the sender's effort as

$$H_S = \log_n m(1 - H_{R|S}).$$

Then we reformulate the objective function, first as

$$\begin{aligned}\Omega_\lambda(H_S) &= \lambda\left(1 - \frac{1}{\log_n m} H_S\right) + (1-\lambda)H_S \\ &= \lambda + \left(1 - \lambda\frac{\log_n m + 1}{\log_n m}\right) H_S\end{aligned}$$

and, second, as

$$\begin{aligned}\Omega_\lambda(H_{R|S}) &= \lambda H_{R|S} + (1-\lambda)\log_n m(1 - H_{R|S}) \\ &= (1-\lambda)\log_n m + (\lambda(1+\log_n m) - \log_n m)H_{R|S}.\end{aligned}$$

If $\lambda < \log_n m/(\log_n m + 1)$, then the slope of the function $\Omega_\lambda(H_S)$, linear in terms of H_S, is positive, and its minimum is attained at the lower boundary $H_S = 0$. At the same time, if $\lambda < \log_n m/(\log_n m + 1)$, then the slope of the function $\Omega_\lambda(H_{R|S})$, linear in terms of $H_{R|S}$, is negative, and its minimum is attained at the upper boundary $H_{R|S} = H_{\max}$. These two conditions yield

$$p(s_i, r_j) = \begin{cases} \dfrac{1}{m} & \text{for a particular } i^* \text{ and all } j; \\ 0 & i \neq i^* \text{ and all } j. \end{cases}$$

If $\lambda > \log_n m/(\log_n m + 1)$, then the slope of the function $\Omega_\lambda(H_S)$, linear in terms of H_S, is negative, and its minimum is attained at the upper boundary $H_S = H_{\max}$. At the same time, if $\lambda > \log_n m/(\log_n m + 1)$, then the slope of the function $\Omega_\lambda(H_{R|S})$, linear in terms of $H_{R|S}$, is positive, and its minimum is attained at the lower boundary $H_{R|S} = 0$. These two conditions yield that there is a function $f : R \to S$, such that

$$p(s_i, r_j) = \begin{cases} \dfrac{1}{m} & \text{for } i^* \text{ and } j^* \text{ where } s_i^* = f(r_j^*); \\ 0 & \text{otherwise,} \end{cases}$$

and such that

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) = \frac{1}{n}.$$

If $\lambda = \log_n m/(\log_n m + 1)$, then the function Ω_λ does not depend on H_S and $H_{R|S}$, and any p attains its minimum. □

Appendix D

Lemma. *The following condition is necessary for $P(\rho) \propto 1/\rho$, where ρ is the rank of the frequency distribution, for $n \to \infty$, assuming that $\log_n m$ is finite:*

$$\frac{2\log_n m - 1}{\log_n m} H_S = H_{R|S}. \tag{D.1}$$

Proof. Let us assume $P(\rho) \propto 1/\rho$, where ρ is the rank of the frequency distribution. The probability $P(\rho)$ must satisfy $\sum_{\rho=1}^{n} P(\rho) = 1$, so

$$\frac{1}{A_n} \sum_{\rho=1}^{n} \frac{1}{\rho} = 1 \tag{D.2}$$

for some constant A_n that depends on n only. Hence,

$$A_n = \sum_{\rho=1}^{n} \frac{1}{\rho} \tag{D.3}$$

is the nth harmonic number that can also be expressed analytically as $A_n = \gamma + \psi_0(n+1)$, where γ is the Euler–Mascheroni constant $(0.577\,215\,6649\cdots)$ and $\Psi(x) = \psi_0(x)$ is the digamma function. Asymptotically,

$$\lim_{n\to\infty} A_n = \ln n + \gamma. \tag{D.4}$$

Substituting $P(\rho) = 1/A_n\rho$ into (1) yields

$$H_S = -\sum_{\rho=1}^{n} \frac{1}{A_n\rho} \log_n \frac{1}{A_n\rho}, \tag{D.5}$$

producing [17]

$$H_S = -\frac{A_n \ln(1/A_n) + \gamma_1(n+1) - \gamma_1}{A_n \ln n}, \tag{D.6}$$

where $\gamma_1 = -0.072\,815\,845\,48\cdots$ is a Stieltjes constant[6], and $\gamma_1(n+1)$ is the generalized Stieltjes constant[7]. The following term converges to zero:

$$\lim_{n\to\infty} -\frac{A_n \ln(1/A_n)}{A_n \ln n} = \lim_{n\to\infty} \frac{\ln A_n}{\ln n} = \lim_{n\to\infty} \frac{\ln(\ln n + \gamma)}{\ln n} = 0,$$

where expression (D.4) is used at the second-last step. This leaves only the term

$$\lim_{n\to\infty} H_S = \lim_{n\to\infty} -\frac{\gamma_1(n+1)}{A_n \ln n}. \tag{D.7}$$

[6] Stieltjes constants are coefficients in the Laurent expansion of the Riemann zeta function $\zeta(z)$ about $z = 1$, given by $\gamma_n = \lim_{M\to\infty} \sum_{i=1}^{M} ((\ln i)^n/i) - ((\ln M)^{n+1}/(n+1))$ [18].

[7] The generalized Stieltjes constant $\gamma_1(a)$ is the first coefficient in the Laurent expansion of the Hurwitz zeta function $\zeta(s, a)$ about $s = 1$ [19].

Connon [20] noted that

$$\lim_{u\to\infty}\left[\gamma_1(u)+\tfrac{1}{2}\ln^2(u)\right]=0.$$

This resolves the remaining term (D.7) as

$$\lim_{n\to\infty}H_S=\lim_{n\to\infty}\frac{1}{2}\frac{\ln^2(n+1)}{A_n\ln n}=\frac{1}{2}\lim_{n\to\infty}\frac{\ln^2(n+1)}{(\ln n+\gamma)\ln n},$$

resulting in

$$\lim_{n\to\infty}H_S=\tfrac{1}{2}. \tag{D.8}$$

Using lemma 2, we obtain

$$\lim_{n\to\infty}H_{R|S}=\lim_{n\to\infty}\left[1-\frac{1}{\log_n m}H_S\right]=\lim_{n\to\infty}\frac{2\log_n m-1}{2\log_n m}.$$

Hence, as $n\to\infty$,

$$\frac{H_{R|S}}{H_S}=\frac{2\log_n m-1}{\log_n m},$$

immediately producing the lemma. □

Appendix E

Lemma. *The number of matrices $\mathcal{L}$ described by the configuration vector $\pi=(\pi_0\cdots\pi_k\cdots\pi_m)$, where π_k are non-negative integers, $0\le k\le m$, is*

$$\mathcal{L}(\pi)=\frac{n!m!}{\prod_{k=0}^{m}\pi_k!(k!)^{\pi_k}}. \tag{E.1}$$

Proof. To reiterate, the overall number of different instances for a configuration is simply

$$\mathcal{L}(\pi)=\mathcal{L}_s(\pi)\cdot\mathcal{L}_r(\pi),$$

where

$$\mathcal{L}_s(\pi)=\frac{n!}{\prod_{k=0}^{m}\pi_k!} \tag{E.2}$$

is the number of permutations of matrix rows, and

$$\mathcal{L}_r(\pi)=\frac{m!}{\prod_{k=0}^{m}(k!)^{\pi_k}} \tag{E.3}$$

captures the number of possibilities to permute 'ones' across an individual row, that is, to permute across m columns.

For example, the configuration $(2;3;2;0^{[5]})$ for a 7×7 matrix has two rows with zeros ($\pi_0=2$), three rows with a single 'one' ($\pi_1=3$), and two rows with two 'ones' ($\pi_2=2$). That is, there are three distinct 'letters' ('zeros', 'singles' and 'doubles') to permute in a seven-letter word. It is well known that the expression (E.2) gives the number of all

possible permuted words, and in this example there are $\mathcal{L}_s(2;3;2;0^{[5]}) = 7!/(2!3!2!) = 210$ possibilities.

Let us consider the term $\mathcal{L}_r(\pi)$. This term captures the number of possibilities to permute 'ones' across an individual row, that is, to permute across m columns. For example, let us consider 'singles'—the rows with a single 'one'. The first of those has $\binom{m}{1}$ choices, the second has $\binom{m-1}{1}$ choices, and the third one has $\binom{m-(1+1)}{1}$ choices—that is, the overall number of permutations for 'singles' is $\mathcal{L}_r(1) = \binom{7}{1}\binom{6}{1}\binom{5}{1} = 210$. Abbreviating

$$m_k = m - \sum_{j=1}^{k-1} j\pi_j$$

for $0 < k \leq m$, it can be easily seen that the number $\mathcal{L}_r(k)$ of choices to permute columns in π_k rows with k 'ones' is given by

$$\begin{aligned}\mathcal{L}_r(k) &= \binom{m_k}{k}\binom{m_k-k}{k}\cdots\binom{m_k-\overbrace{(k+\cdots+k)}^{k(\pi_k-1)}}{k} \\ &= \frac{m_k!}{k!(m_k-k)!}\frac{(m_k-k)!}{k!((m_k-k)-k)!}\cdots\frac{(m_k-k(\pi_k-1))!}{k!(m_k-k\pi_k)!} \\ &= \frac{m_k!}{\underbrace{k!\cdots k!}_{\pi_k}(m_k-k\pi_k)!}.\end{aligned}$$

That is,

$$\mathcal{L}_r(k) = \frac{m_k!}{(k!)^{\pi_k}(m_k-k\pi_k)!}.$$

In our example, $\mathcal{L}_r(1) = 7!/((1!)^3(7-3)!) = 210$, and $\mathcal{L}_r(2) = (7-3)!/((2!)^2((7-3)-2\cdot 2)!) = 6$.

In order to produce the total number, one simply needs to multiply the terms $\mathcal{L}_r(k)$, each of which further reduces the total number of choices m by the number of 'ones' already dealt with:

$$\mathcal{L}_r(\pi) = \prod_{k=1}^{m}\mathcal{L}_r(k) = \prod_{k=1}^{m}\frac{m_k!}{(k!)^{\pi_k}(m_k-k\pi_k)!}.$$

Noticing that $m_1 = m$, $m_{k+1} = m_k - k\pi_k$ for $k > 0$, and $m_k = m\pi_m$ for $k = m$ (so that $(m_k - m\pi_m)! = 1$ for $k = m$), this can be further reduced as follows:

$$\mathcal{L}_r(\pi) = \frac{m!}{\prod_{k=1}^{m}(k!)^{\pi_k}} = \frac{m!}{\prod_{k=0}^{m}(k!)^{\pi_k}},$$

immediately yielding expression (E.3) and the lemma. □

For example, $\mathcal{L}_r(2;3;2;0^{[5]}) = 7!/(0!^2\cdot 1!^3\cdot 2!^2) = 1260$. The overall number of different instances for our example configuration $(2;3;2;0^{[5]})$ is given by $\mathcal{L}(\pi) = \mathcal{L}_s(\pi)\cdot\mathcal{L}_r(\pi) = 210\cdot 1260 = 264\,600$. It turns out to be the largest number of instances across all configurations for a 7×7 minimizer, amounting to over 32% of all the instances.

Appendix F

Lemma. [8] *The vector* $\pi^* = (\pi_0^* \cdots \pi_k^* \cdots \pi_m^*)$, *where* π_k^* *are real numbers, that maximizes* $\mathcal{L}(\pi)$ *under the conditions* $\sum_{k=0}^m \pi_k = n$ *and* $\sum_{k=1}^m k\pi_k = m$, *asymptotically follows the Poisson distribution with the average* (m/n) *being multiplied by* n*:*

$$\pi_k^* \approx \frac{n\mathrm{e}^{-(m/n)}(m/n)^k}{k!}. \tag{F.1}$$

Proof. In order to maximize $\mathcal{L}(\pi)$, we need to minimize $\prod_{k=0}^m \pi_k!(k!)^{\pi_k}$. Since the conditions are linear in π, we minimize the logarithm:

$$\ln\left(\prod_{k=0}^m \pi_k! \prod_{k=0}^m (k!)^{\pi_k}\right) = \sum_{k=0}^m \ln \pi_k! + \sum_{k=0}^m \pi_k \ln(k!).$$

Differentiating over π, and using the Gamma function (extension of the factorial function), i.e., $\pi_k! = \Gamma(\pi_k + 1)$, yields the condition for the optimal π_k^*:

$$\frac{\mathrm{d}\ln\Gamma(\pi_k^* + 1)}{\mathrm{d}\pi_k^*} + \ln(k!) = \mu_1' + k\mu_2',$$

where μ_1' and μ_2' are Lagrange multipliers for the conditions. Using the digamma function Ψ, that is, the logarithmic derivative of the Gamma function $\Psi(\pi_k^* + 1) = \mathrm{d}\ln\Gamma(\pi_k^* + 1)/(\mathrm{d}\pi_k^*)$, we obtain

$$\Psi(\pi_k^* + 1) + \ln(k!) = \ln \mu_1 + k \ln \mu_2,$$

where $\mu_1 = \mathrm{e}^{\mu_1'}$ and $\mu_2 = \mathrm{e}^{\mu_2'}$. For large arguments $\Psi(x + 1) \approx \ln(x)$, and within this approximation

$$\pi_k^* \approx \frac{\mu_1 \mu_2^k}{k!}. \tag{F.2}$$

The condition $\sum_{k=0}^m \pi_k = n$ yields

$$n \approx \sum_{k=0}^m \frac{\mu_1 \mu_2^k}{k!} = \frac{\mu_1 \mathrm{e}^{\mu_2}\Gamma(m + 1, \mu_2)}{m!} \approx \mu_1 \mathrm{e}^{\mu_2}.$$

The condition $\sum_{k=1}^m k\pi_k = m$ produces

$$m \approx \sum_{k=1}^m k\frac{\mu_1 \mu_2^k}{k!} = \frac{\mu_1 \mu_2 \mathrm{e}^{\mu_2}\Gamma(m, \mu_2)}{(m-1)!} \approx \mu_1 \mu_2 \mathrm{e}^{\mu_2}.$$

Using the last two equations we obtain

$$\mu_2 \approx \frac{m}{n} \tag{F.3}$$

and

$$\mu_1 \approx n\mathrm{e}^{-\mu_2} = n\mathrm{e}^{-(m/n)}. \tag{F.4}$$

[8] The authors thank an anonymous referee for pointing out the Poisson distribution as the optimal solution, and for the proof.

Substituting (F.3) and (F.4) into (F.2) yields

$$\pi_k^* \approx \frac{n\mathrm{e}^{-(m/n)}(m/n)^k}{k!}. \tag{F.5}$$

That is, the optimal vector π^* follows the Poisson distribution with the average (m/n) being multiplied by n. □

Appendix G

The rank ρ_k of the sequence of signals s_i sorted by their frequency $P(\rho_k) \equiv (k/m)$ satisfies the following condition:

$$m - \sum_{j=0}^{k} \pi_j < \rho_k \leq m - \sum_{j=0}^{k-1} \pi_j. \tag{G.1}$$

For example, the rank of signals which encode one object (i.e., $k = 1$) satisfies

$$m - \pi_0 - \pi_1 < \rho_1 \leq m - \pi_0.$$

Let us consider the lower bound, using the solution (20):

$$\rho_k = m - \sum_{j=0}^{k} \pi_j = m - \sum_{j=0}^{k} \frac{m}{\mathrm{e}j!} = m - \frac{m}{\mathrm{e}} \sum_{j=0}^{k} \frac{1}{j!}.$$

Taylor expansion for the exponential function e^x, at $x = 1$, yields

$$\sum_{j=0}^{k} \frac{1}{j!} = \mathrm{e} - R_k(1),$$

where $R_k(x)$ is the remainder term of the kth order Taylor approximation to e^x:

$$R_k(1) = \frac{\mathrm{e}^{\xi}}{(k+1)!}$$

for a number ξ between 0 and 1. In fact, $0 < \xi \leq \ln(\mathrm{e} - 1)$. Hence, the lower bound is given by

$$\rho_k = m - \frac{m}{\mathrm{e}} \left(\mathrm{e} - \frac{\mathrm{e}^{\xi}}{(k+1)!} \right) = \frac{m\mathrm{e}^{\xi-1}}{(k+1)!}.$$

Substituting $(k-1)$ for k gives the upper bound of expression (G.1), establishing

$$\frac{m\mathrm{e}^{\xi-1}}{(k+1)!} < \rho_k \leq \frac{m\mathrm{e}^{\xi-1}}{k!}. \tag{G.2}$$

Expressing $(k+1)!$ as a function of the rank ρ_k leads to the approximation

$$(k+1)! \approx \frac{m\mathrm{e}^{\xi-1}}{\rho_k}$$

or

$$\Gamma(k) \approx \frac{m\mathrm{e}^{\xi-1}}{\rho_k}.$$

Cantrell [16] noted that for $x \geq x_0$, where x_0 denotes the positive zero of the digamma function ($x_0 \approx 1.461\,632$), $\Gamma(x)$ is strictly increasing. Hence, restricting its domain accordingly, the inverse is a function given by

$$\Gamma^{-1}(x) = \frac{L(x)}{W(L(x)/\mathrm{e})} + 1/2 \tag{G.3}$$

where

$$L(x) \approx \ln \frac{x + 0.036\,534}{\sqrt{2\pi}}$$

and $W(z)$ is the principal branch of the Lambert W function (the product logarithm), i.e., the inverse function of the function $f(w) = w\mathrm{e}^w$. Using the inverse function $\Gamma^{-1}(x)$ we obtain

$$k \approx \Gamma^{-1}\left(\frac{m\mathrm{e}^{\xi-1}}{\rho_k}\right).$$

Setting $P(\rho_k) = (k/m)$ yields

$$P(\rho_k) \approx \frac{1}{m}\Gamma^{-1}\left(\frac{m\mathrm{e}^{\xi-1}}{\rho_k}\right).$$

It is clear that this dependency reduces the symbols' frequency much slower than a power law—in fact, the rate of change is sub-logarithmic.

References

[1] Ferrer i Cancho R and Elvevg B, *Random texts do not exhibit the real Zipf's law-like rank distribution*, 2010 *PLoS ONE* **5** e9411
[2] Manin D Y, *Zipf's law and avoidance of excessive synonymy*, 2008 *Cogn. Sci.: a Multidiscip. J.* **32** 1075
[3] Ferrer i Cancho R, *The variation of Zipf's law in human language*, 2005 *Eur. Phys. J.* B **44** 249
[4] Newman M E J, *Power laws, Pareto distributions and Zipf's law*, 2005 *Contemp. Phys.* **46** 323
[5] Gabaix X, *Zipf's law for cities: an explanation*, 1999 *Q. J. Econom.* **114** 114
[6] Ferrer i Cancho R and Solé R V, *Least effort and the origins of scaling in human language*, 2003 *Proc. Nat. Acad. Sci.* **100** 788
[7] Ferrer i Cancho R, *Zipf's law from a communicative phase transition*, 2005 *Eur. Phys. J.* B **47** 449
[8] Ferrer i Cancho R and Díaz-Guilera A, *The global minima of the communicative energy of natural communication systems*, 2007 *J. Stat. Mech.* P06009
[9] Kosmidis K, Kalampokis A and Argyrakis P, *Statistical mechanical approach to human language*, 2006 *Physica* A **366** 495
[10] Hunt A, *Relevance of percolation theory to power-law behavior of dynamic processes including transport in disordered media*, 2009 *Complexity* **15** 13
[11] Tamaoka K, Meyer P, Makioka S and Altmann G, *On the dynamics of the compounding of Japanese kanji with common and proper nouns*, 2008 *J. Quantitative Linguistics* **15** 136
[12] von Ferber C, Holovatch T, Holovatch Y and Palchykov V, *Modeling Metropolis Public Transport*, 2009 *Traffic and Granular Flow'07* ed C Appert-Rolland, F Chevoir, P Gondret, S Lassarre, J P Lebacque and M Schreckenberg (Berlin: Springer) pp 709–19
[13] Tieri P, Valensin S, Franceschi C, Morandi C and Castellani G C, *Memory and selectivity in evolving scale-free immune networks*, 2003 *Artificial Immune Systems, Proceedings* (*Springer Lecture Notes in Computer Science* vol 2787) ed J Timmis, P Bentley and E Hart (Berlin: Springer) pp 93–101
[14] Obst O, Prokopenko M and Polani D, *Origins of scaling in genetic code*, 2009 *Proc. European Conf. on Artificial Life (ECAL)* (*Lecture Notes in Computer Science* vol 5777–5778) Budapest at press
[15] Trosso A, *La legge di Zipf ed il principio del minimo sforzo*, 2008 *Zipf's law and the least effort principle* March–April 2008, University of Turin, Italy

J. Stat. Mech. (2010) P11025

[16] Cantrell D W, *Inverse gamma function*, 2001
http://mathforum.org/kb/message.jspa?messageID=342551&tstart=0
[17] Wolfram Alpha L L C, *Wolfram|Alpha*, 2010
http://www.wolframalpha.com/input/?i=-sum_(i=1)^n+(1/(A+i)+*+log+(1/(A+i))
[18] Havil J, 2003 *Gamma: Exploring Euler's Constant* (Princeton, NJ: Princeton University Press)
[19] Weisstein E W, *Stieltjes constants*, 2010 *From MathWorld–A Wolfram Web Resource*
http://mathworld.wolfram.com/StieltjesConstants.html
[20] Connon D F, *Some applications of the Stieltjes constants*, 2009 arXiv:0901.2083v1

J. Stat. Mech. (2010) P11025