

Research



Cite this article: Villegas P, Muñoz MA, Bonachela JA. 2020 Evolution in the Debian GNU/Linux software network: analogies and differences with gene regulatory networks.

J. R. Soc. Interface **17**: 20190845.

<http://dx.doi.org/10.1098/rsif.2019.0845>

Received: 11 December 2019

Accepted: 21 January 2020

Subject Category:

Life Sciences—Physics interface

Subject Areas:

evolution, biocomplexity, computational biology

Keywords:

evolving networks, network resilience, information transmission

Author for correspondence:

Juan A. Bonachela

e-mail: juan.bonachela@rutgers.edu

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.4840281>.

Evolution in the Debian GNU/Linux software network: analogies and differences with gene regulatory networks

Pablo Villegas^{1,2}, Miguel A. Muñoz¹ and Juan A. Bonachela^{3,4}

¹Departamento de Electromagnetismo y Física de la Materia, Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, 18071 Granada, Spain

²Istituto dei Sistemi Complessi, CNR, via dei Taurini 19, 00185 Rome, Italy

³Marine Population Modeling Group, Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, UK

⁴Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA

JAB, 0000-0002-3316-8120

Biological networks exhibit intricate architectures deemed to be crucial for their functionality. In particular, gene regulatory networks, which play a key role in information processing in the cell, display non-trivial architectural features such as scale-free degree distributions, high modularity and low average distance between connected genes. Such networks result from complex evolutionary and adaptive processes difficult to track down empirically. On the other hand, there exists detailed information on the developmental (or evolutionary) stages of open-software networks that result from self-organized growth across versions. Here, we study the evolution of the Debian GNU/Linux software network, focusing on the changes of key structural and statistical features over time. Our results show that evolution has led to a network structure in which the out-degree distribution is scale-free and the in-degree distribution is a stretched exponential. In addition, while modularity, directionality of information flow, and average distance between elements grew, vulnerability decreased over time. These features resemble closely those currently shown by gene regulatory networks, suggesting the existence of common adaptive pathways for the architectural design of information-processing networks. Differences in other hierarchical aspects point to system-specific solutions to similar evolutionary challenges.

1. Introduction

Understanding the collective properties stemming from the interactions of a large number of units such as genes, proteins or metabolites is of paramount importance in biology [1–3]. Theoretical work focusing on the changes over time of self-organizing networks can provide key information about these natural systems. Particularly, network theory provides us with a highly insightful systems-level perspective to extremely complicated biological problems, which has helped advance knowledge in fields such as neuroscience [4], ecology [5] and epidemiology [6], to name a few [7]. The study of information processing in living systems has greatly benefited from this network perspective, complementing parallel endeavours for the analysis of single pathways, and providing a much richer understanding of collective phenomena emerging from a large number of basic inter-related units [8,9].

More specifically, analyses of gene-regulatory, protein–protein and metabolic networks have led to dramatic advances in systems biology [10,11]. Indeed, an important step forward in the understanding of cell regulatory mechanisms was the discovery of the scale-free degree distribution of such networks [11]. Most networks within the cell show other non-trivial structural features such as a

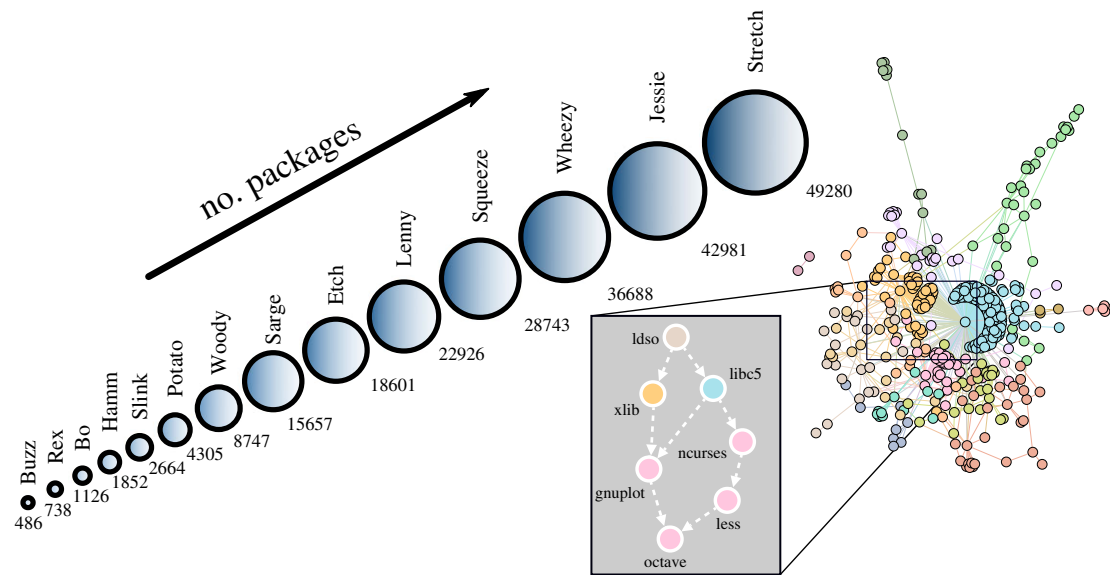


Figure 1. Left: sketch showing the official code-name and the total number of packages of all stable Debian releases (releases named after characters from the movie *Toy Story*[®]); circle sizes are proportional to the number of packages in each distribution. Right: sample of the dependencies network for the first release, where different colours indicate package moduli in the network (which we do not specify here); inset: subgraph showing specific dependencies between some well-known packages from this distribution.

small-world property (arising from a small average path length, or distance between any two elements), a highly modular structure, and an extraordinary responsiveness to environmental cues, as well as dynamical and structural robustness [11]. In particular, gene regulatory networks (GRNs)—which display mutual regulatory interactions between genes—have been identified as directed networks that exhibit a scale-free distribution for the number of regulated genes (called ‘out degree’ due to the direction of information flow) but an exponential distribution for the number of controlling genes (in degree). In addition, cellular states can be identified as attractors of the dynamics of genetic regulatory systems, which allows the latter to be modelled—at least in first approximation—as random Boolean networks [12–15]. In such a simplified approach, mutual regulatory interactions, described as direct links between genes/nodes, involve arbitrary random Boolean functions whose inputs are the on/off states encoding the expression level of other genes. This type of binary set-up has shed light on important conceptual problems such as the emergence of diverse phenotypes from a unique genetic network, the existence of transitions among them (e.g. cell differentiation and reprogramming), and the emergence of cycles in cell states. In this regard, two examples are the predicted expression patterns of the fly *Drosophila melanogaster* [16] and the yeast cell cycle [17]. In either case, it seems clear that GRNs involve some type of information transmission (or flow) encoded in mutual regulatory interactions, determining the cellular response to different stimuli or environmental conditions [11,18]. The analysis and study of interacting systems with similar information flow can help understand the particular structure and emergent properties of genetic systems.

One particular aspect of GRNs that remains elusive is how such properties and functionality have emerged through evolution. The main reason for this knowledge gap is that the information currently available provides a limited picture of the evolutionary path followed by biological networks. Artificial, self-organized networks, on the other hand, can offer an unrivaled level of detail regarding their different ‘evolutionary’ stages. In this respect, self-organized software

networks have been shown to constitute an excellent model for the study of the evolution of biological networks [19–24].

Software networks are composed of packages, acting as nodes and forming the basic unit of software. These are inter-related due to the need for a package to reuse code of other packages in order to work properly (the so-called dependencies, i.e. package *i* needs other packages to be functional). Similar to GRNs, the storage and transmission of information through this network enables the proper functioning of the operating system as a whole. Thus, it is pertinent to ask: what can the characterization of the evolutionary history of a software network reveal about the evolution and main architectural features of GRNs?

To shed light onto this question, we focus here on the Debian GNU/Linux operating system (Debian, hereon), for which publicly available historical data exist. First, we characterize the evolving structure of the network of dependencies between software packages in the different Debian distributions released up to June 2019. Then, we explore the emergent properties of such networks and their role in the functionality of the system. With this information, we finally scrutinize the similarities and differences between the structure and emergent properties of software networks and GRNs.

2. Debian networks

Debian is an open-source operating system that, in the last 25 years, has changed in a self-organized way through the collective action of a myriad of developers. The history of Debian shows many small intermediate steps but only 14 important stable releases (evolutionary steps) that have progressively altered its networked structure. This evolution has resulted in a sustained growth generating an actual network of interactions between thousands of packages, starting from a very small initial release in its first version (figure 1). Such packages must translate the information coming from other packages in order to satisfy the so-called dependencies, i.e. pieces of software required for the package to work. Thus, we can

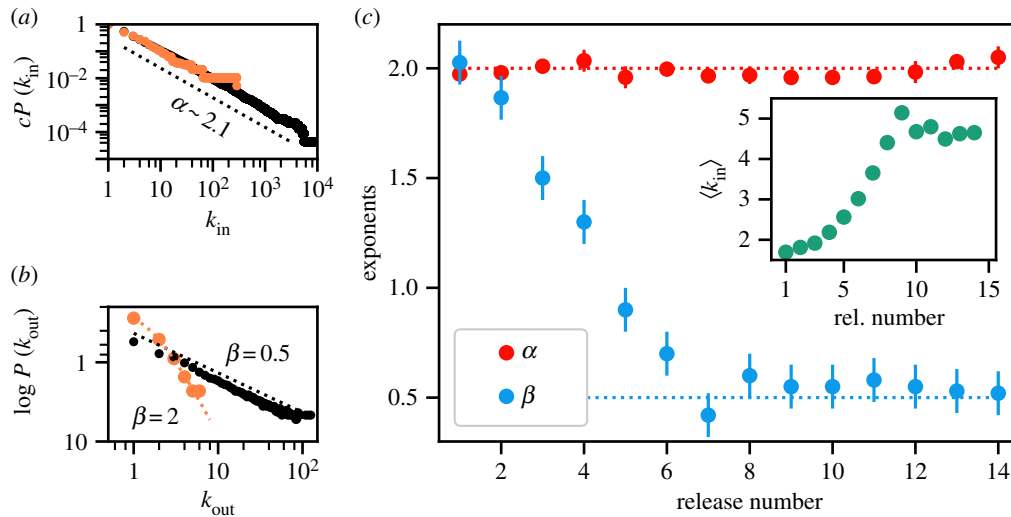


Figure 2. (a) Cumulative degree distribution for outgoing dependencies for the first (orange) and last (black) releases in log–log scale. (b) Logarithm of the degree distribution for in-degree for the same two releases, showing a half-normal distribution decay ($\beta = 2$) for the first version and a stretched exponential ($\beta = 0.5$) for the last one. (c) Exponents from the out-degree distribution of dependencies power law (blue) and the in-degree distribution stretched exponential (red) for the first release to the most recent one. Inset: mean in-degree for all distributions.

represent each version of the operating system using a different network for each of the different evolutionary stages. Nodes represent packages and links join every node with the packages it depends upon, thus forming a directed network of dependencies [20]. Note that the specific interpretation of the flow of information determines the direction of the links. Flow in terms of dependencies imposes that directional links start on a focal package and point to its dependencies (e.g. [20]). Flow in terms of information transmission, like in GRNs, is typically represented with opposite directional links: because i depends on j , information flows from j to i . Here, we use the latter convention to facilitate comparison between the Debian networks and GRNs.

We have focused on the relationships between the binary x86 packages included in Debian, for its first 14 (stable) distributions (figure 1, left). Figure 1 (right panel) shows one specific example of package inter-dependence extracted from the very first distribution. In addition to dependencies reflecting requirements, additional relationships between packages reflect incompatibilities among them (i.e. ‘conflicts’) [25]. In most cases, conflicts are used to avoid duplication in virtual facilities of the system, as explained in the Debian Policy Manual [26]. Most conflicts occur between potential candidates to fulfil a requirement or function, grouped under the term ‘virtual packages’. Such virtual facilities are supplied by particular choices among many possibilities (e.g. both *firefox* and *konqueror* provide the same service, *www-browser*). Here, we resolve conflicts by averaging over the ensemble of potential networks resulting from random choices for each ‘virtual package’ (see Methods), which allows us to focus only on the network of mutual dependencies.

3. Results

Since the initial Buzz distribution, which included 486 packages, the total number of packages forming each Debian stable version has grown exponentially (see figure 1 and electronic supplementary material, SI-1). Here, we scrutinize this growth across releases (evolutionary steps), and analyse how it translated into the emergent structural properties of the associated Debian software networks.

3.1. Degree distributions

To characterize the changes in the structural properties of Debian over time, we first analysed the in- and out-degree distributions (averaged over realizations of the network obtained after resolving conflicts). The former describes the probability for a package to depend on k_{in} packages, whereas the latter describes the probability for a package to be a requirement for k_{out} packages. As shown in figure 2a and 2b (respectively), the cumulative out-degree distribution exhibits a power law tail, $P(k_{\text{out}}) \sim k_{\text{out}}^{-\alpha+1}$, with characteristic exponent α , and the in-degree distribution follows a stretched exponential, $P(k_{\text{in}}) \sim \exp(-(k_{\text{in}}/\tau)^\beta)$, with characteristic exponent β (see Methods and electronic supplementary material, SI-2, for fits and estimations of their likelihood). As shown in figure 2c, the power law for the out-degree distribution shows an exponent very close to $\alpha = 2$ for all releases (red points). The exponent of the stretched exponential for the in-degree distributions (blue points) decays from $\beta = 2$ in the early releases (normal distribution), to an approximately stationary value $\beta \approx 0.5$ after the eighth release. The inset of figure 2c shows that the mean in-degree connectivity, $\langle k_{\text{in}} \rangle$, grows from 1.5 to an apparently stationary value approximately 4.75, reached after the eighth release.

3.2. Emergence of a non-trivial modular structure

We also quantified computationally how modular is the structure of each release’s network, for which we used Newman’s modularity index [7] (Q , see Methods). As figure 3a shows, modularity grows continuously for the last six versions (i.e. after the eighth release), from 0.5 to approximately 0.7 (orange points). We further explored the possibility for such modularity to be an artefact of the degree sequence of the nodes. To this end, and although Newman’s modularity index already discounts the random expectation for random networks, we constructed null/random structure networks by randomizing the original networks exchanging links between random nodes under certain rules. Specifically, for each original (conflict-resolved) Debian network (D), we constructed a random copy (D') in which a random swapping process maintained both the incoming and the outgoing connectivity of each single node unaltered

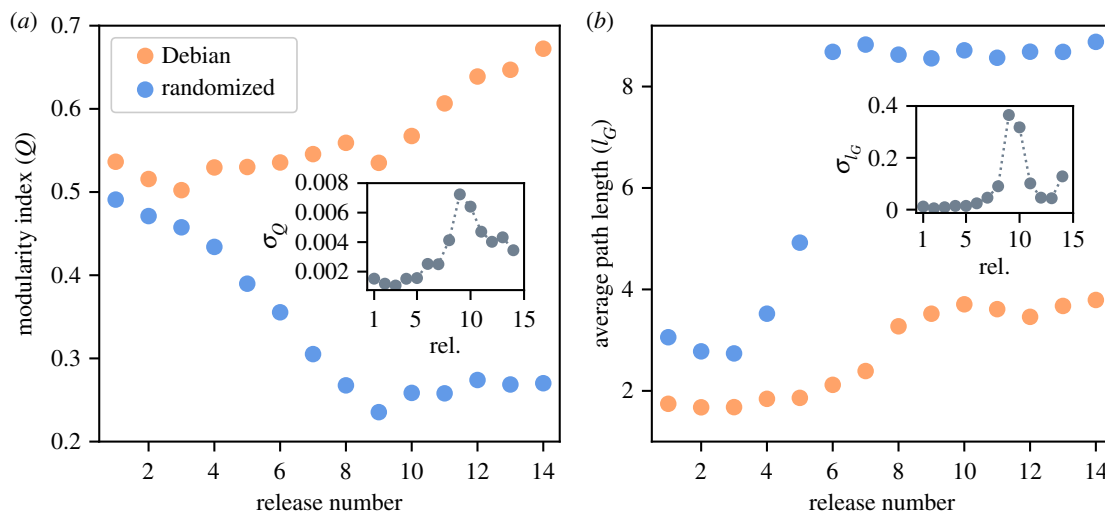


Figure 3. (a) Modularity index (Q) for the different Debian releases (orange points) and the ensemble of random (swapped) networks (blue points). (b) Average path length (l_G) for Debian networks (orange points) and the ‘swapped’ networks (blue points). In both panels, insets show the standard deviation for the ‘swapped’ networks. Q and l_G are averaged over 10^4 realizations.

(see Methods). As illustrated by the blue points in figure 3a, the randomized networks show a sustained decrease of their modularity index until the eighth distribution. The standard deviation of the modularity index for the ‘swapped’ networks shows a pronounced peak around such version as well (see inset). Moreover, the randomization allowed us to calculate a Z-score for Q to estimate the difference in modularity between the Debian networks and their randomized counterparts (normalized by the available diversity, i.e. standard deviation, of the ensemble of possible networks; see Methods). This Z-score shows a steep increase over time, i.e. the modularity of Debian becomes progressively larger than that of its random-network counterparts over time (see electronic supplementary material, figure S4).

We also measured how the average path length (l_G) changed across the different Debian releases. Figure 3b shows a small increase in l_G for the original networks (orange points) constrained in (1.75–3.8). This growth slows down after version 8. For the ‘swapped’ networks (blue points), l_G shows a steep increase that jumps to a considerably higher value (approx. 8.75) around the sixth version, and remains at that level afterwards. This gap approximately coincides with a pronounced peak for its standard deviation in the ensemble of randomized networks (see inset). The Z-score for the average path length grows monotonically, showing two growth regimes before and after the eighth release (see electronic supplementary material, figure S4). As the size of the network increases, the distance that information needs to travel between any two packages in Debian becomes larger on average, but grows much more slowly than that for the randomized counterparts. Moreover, Debian grew accordingly to the expectation for a typical small-world network (i.e. proportionally to the logarithm of the number of elements; see electronic supplementary material, figure S5), with different trends before and after the eighth release.

3.3. Evolving hierarchy

To further analyse the evolution of the network structure, we explored how diverse aspects of the hierarchy of the Debian networks changed with time. Due to the multi-faceted character of hierarchy, there is no single definition or observable

to measure [27], and hence we monitored three different indicators: (a) the level of ‘stratification’ or re-use of software in the network; (b) the directionality of the information flow; and (c) the level of reciprocal dependence between packages (see Methods).

To measure the stratification/re-use of software in the network, we introduced three categories for packages [28]: (i) packages with no dependencies (‘sources’ or information containers); (ii) packages that depend on other packages and, at the same time, some packages depend on them (‘pass-through’ nodes); and (iii) packages that no other package depends upon (‘sinks’). In this way, each network is decomposed into three distinct hierarchical levels or strata. We monitored the fraction of packages within each of these hierarchical categories across the different releases. As shown in figure 4a, ‘source’ packages only represent a small fraction of any given network, whereas the proportion of ‘sinks’ decreases progressively from a remarkable 0.8 value observed for the first distributions. At the same time, the number of ‘pass-through’ nodes increases, indicating that most sinks moved to this category (i.e. packages tend to be re-used).

To quantify the directionality of information flow, we measured the so-called flow index, χ [29]. In a nutshell, χ quantifies the fraction of links from lower to higher hierarchical levels, i.e. fraction of links that are aligned with an inherent directionality. A value of χ close to 1 indicates that there is an overall directional flow such that a large fraction of links point from higher to lower hierarchical levels. As figure 4b shows, Debian overall exhibits high directionality (i.e. low $1 - \chi$ values). Although, for earlier releases, the fraction of links pointing in the direction of the flow decreases discontinuously, the trend inverts after the eighth release, and $1 - \chi$ shows a steep decrease (i.e. increase of directionality) over the following versions of Debian.

To quantify the reciprocal dependence between packages, we measured a classic hierarchy index, the Krackhardt hierarchy score, K_{HS} [30]. Here, we represented $1 - K_{HS}$, which indicates the fraction of pairs that show mutual or reciprocal dependence (i.e. package i can be reached from package j and the other way around). Figure 4c shows a monotonic decrease for $1 - K_{HS}$, representing a loss of symmetrically linked pairs in the network across releases.

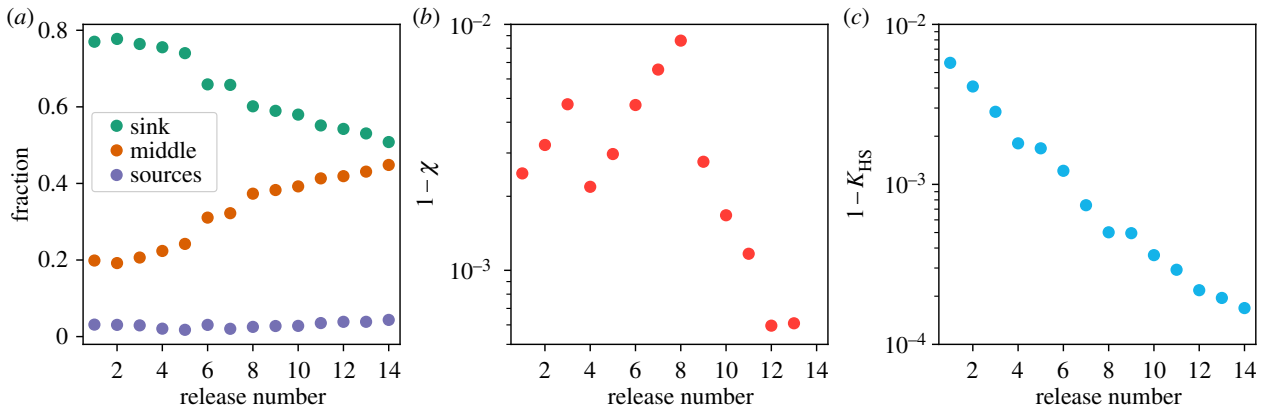


Figure 4. (a) Evolution of the three main hierarchical categories defined in the network: ‘sinks’, ‘pass-through nodes’ and ‘sources’. Observe that the ‘sources’ of information, i.e. nodes that do not depend on any node, constitute a small fraction of the network. However, ‘sinks’, situated in the lowest point of the hierarchy, are disappearing over time. (b) Flow index (represented as $1 - \chi$ to allow for the semi-log plot), showing an initial discontinuous decrease followed by a monotonic increase after the eighth version. (c) Reciprocal of the Krackhardt hierarchy score, $1 - K_{HS}$, which decreases monotonically for all the Debian releases. Altogether points to a dynamic change of the various aspects of hierarchy for the network.

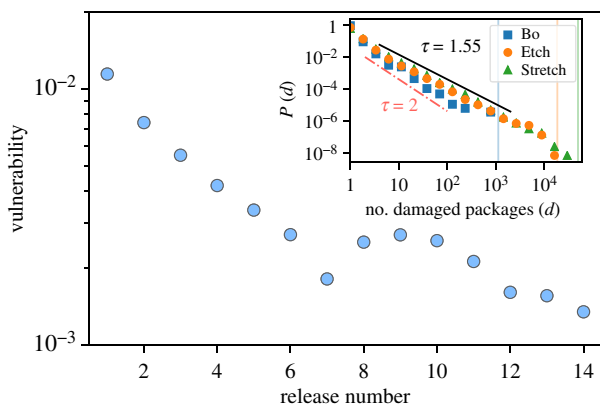


Figure 5. Vulnerability index for the different Debian distributions obtained after 10^4 ‘damage’ experiments, showing an overall decrease of vulnerability. Inset: scale-free avalanches of damage after an attack to one random node of the network for three representative releases (averaged over possible initial seeds and networks in the ensemble). Lines are guides to the eye for the slopes indicated by the exponents (dashed-dotted red line for $\tau = 2$, and black line for $\tau = 1.55$). Vertical lines represent the total number of packages in each distribution, which is indicative of upper cutoffs. Error bars are smaller than the symbols.

In summary, the three measures altogether show that the hierarchy of the Debian networks has changed over time, strengthening the overall directionality of the flow and keeping a low reciprocal dependence, while increasing the re-use of software.

3.4. Evolution of robustness

One plausible expectation for the development of any operating system is that the ability of its associated network to retain functionality should increase over versions in spite of damages, i.e. it is expected to enhance its robustness. To explore this possibility, we perturbed each Debian network by removing at random one package (i.e. we made a package unusable), and followed the consequent cascade of failures by removing nodes depending upon removed packages, which ultimately affects a certain fraction of the whole network. We called this fraction of affected nodes ‘vulnerability’. More precisely, we defined vulnerability as the fraction of affected packages averaged over initial damage ‘seeds’ and network ensemble. Figure 5 shows that the vulnerability

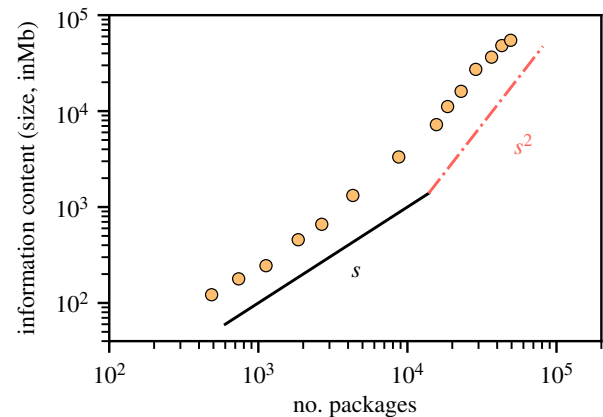


Figure 6. Total number of packages measured as a function of the system size for each Debian release. After the initial, approximately linear growth, a slowing down can be appreciated as general trend. Lines are guides to the eye showing a linear growth (black) and a quadratic growth (red).

index decreases across versions, and exhibits a slow-down for this trend after a peak around the eighth to ninth distributions.

We also analysed the full probability distribution for the number of packages affected by each cascade of failures (or avalanches size, using the jargon from damage spreading analyses [31]). This avalanche-size distribution shows a power-law tail with an exponent that decreases over time (from an initial $\tau = 2$ to $\tau = 1.55$ for the last distribution; inset of figure 5).

3.5. Network information content

We also studied the relationship between the total size (measured in megabytes) of the different packages, as a proxy for the total amount of information contained in the Debian network. In particular, we analysed the evolution of the sum of the sizes of all packages for each potential network as a function of the number of packages (figure 6). As happens with other properties, the information content grows monotonically across time, but it changes its slope around the eighth distribution, moving from a linear to a superlinear trend (i.e. the information content of the network grows faster than the number of packages).

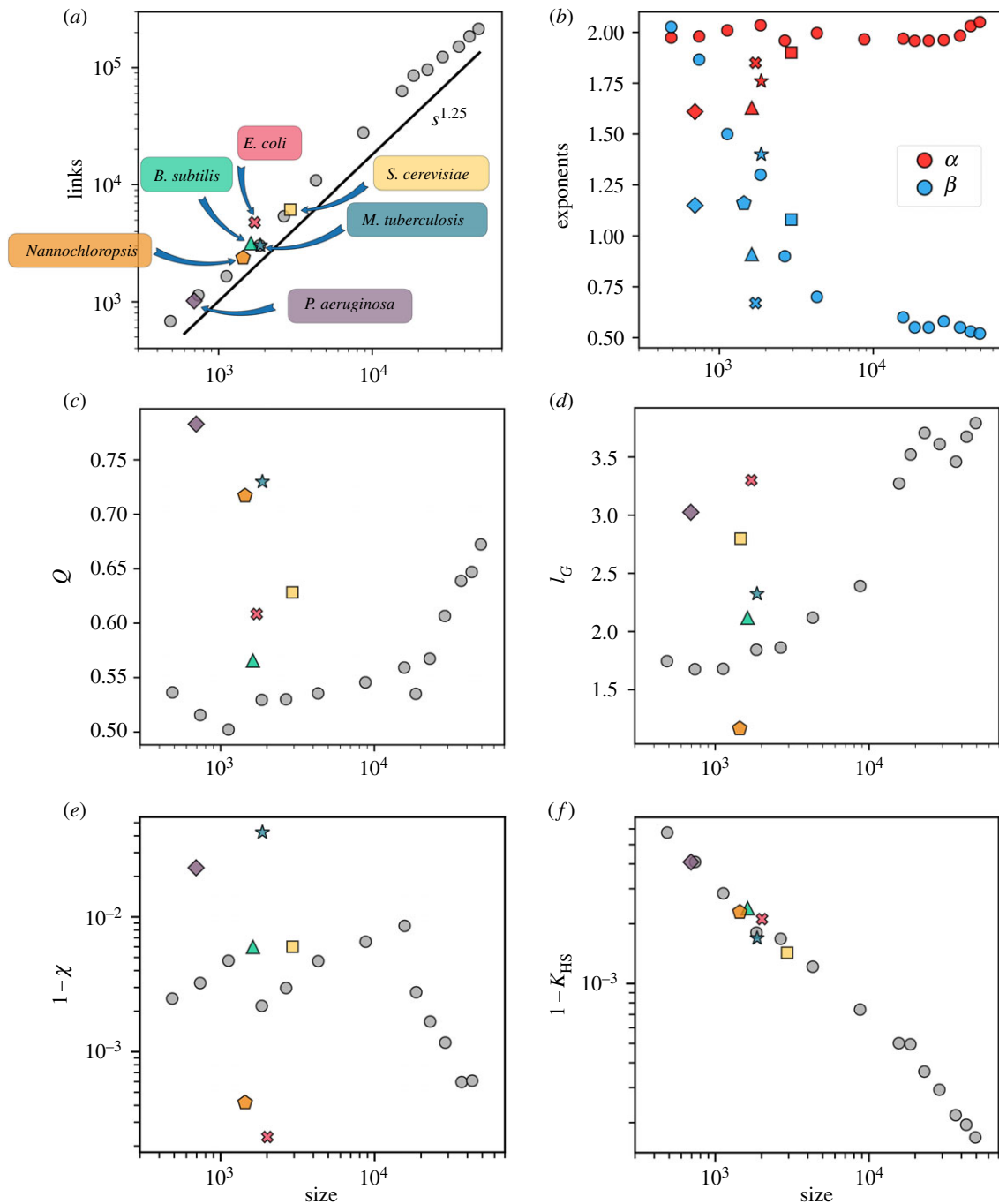


Figure 7. Comparison between Debian and GRNs for different properties dependent on size (horizontal axis for all panels). (a) Total number of links versus system size, with Debian networks represented by grey circles and GRNs by symbols. (b) Exponents associated with the scale-free (α) and stretched exponential (β) degree distributions mentioned in the previous sections. (c) Newman modularity index (Q) for Debian networks (grey circles) and GRNs (symbols). (d) Average path length versus system size for Debian networks (grey circles) and GRNs (symbols). (e) One minus the flow index, $1 - \chi$, for different Debian releases and GRNs as a function of the system size. (f) One minus the Krackhardt hierarchy score, $1 - K_{HS}$.

3.6. Results for gene-regulatory networks

We analysed GRNs from publicly available datasets for *E. coli*, *M. tuberculosis*, *B. subtilis*, *P. aeruginosa* and *S. cerevisiae* [32–35]. Below, we summarize our main findings, and refer the reader to electronic supplementary material, SI-4 and SI-5, for a comprehensive presentation of all these results.

GRNs are very sparse networks, with mean incoming connectivity around 1.5–3.5 in most of the cases [15]. They show a total number of links for a given total number of nodes similar to that of the Debian networks (figure 7a were coloured nodes of diverse shapes correspond to GRNs). The distribution for the in-degree—the number of regulatory interactions in which a given gene intervenes—

has been reported to be exponential [11], whereas the distribution for the out-degree—the number of interactions controlling the expression of a given gene—is scale free [11]. Our own analyses reveal that the out-degree distribution decays as a power law, $P(k_{out}) \sim k_{out}^{-\alpha}$, with characteristic exponent $\alpha \in (1.5 - 2)$ [10,36,37], and the in-degree distribution follows a stretched exponential, $P(k_{in}) \sim \exp(-k_{in}/\tau^\beta)$, with characteristic exponent $\beta \in (0.5 - 1.5)$ (see electronic supplementary material, SI-4). These values are similar to those shown by the Debian networks, given their respective network sizes (figure 7b). For the analysed GRNs, Newman’s modularity index Q is in the interval (0.55–0.8), and the average path length l_G in (2.0–3.5)

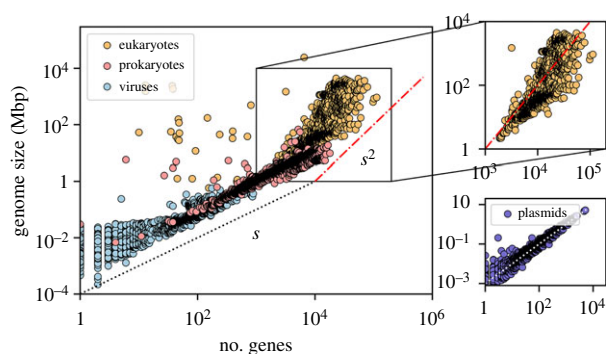


Figure 8. Genome size as a function of the total number of genes (in log–log scale) for multiple microbial organisms (data from the NCBI database [39]). Observe the linear scaling associated with prokaryotes and viruses (red and blue points, respectively) as well as plasmids (lower inset, violet points). On the other hand, data for eukaryotes depart from this linear relationship, becoming steeper to an approximately quadratic scaling (orange dots, enlarged in the upper inset). Lines are guides to the eye.

(figure 7*c,d*), values that are lower than those of the Debian networks. On the other hand, GRNs show a very prominent hierarchical structure: more than 85% of the nodes are ‘sinks’, i.e. they do not control other genes [28,38], there is a very strong directionality of the information flow (i.e. very small $1 - \chi$ value), and the fraction of reciprocally expressed genes, $1 - K_{HS}$, is very low (figure 7*e,f*). As these panels show, GRNs show a range of values for the directionality of information flow that is wider than that shown by the Debian networks, and both systems share similarly small fractions of information-unit pairs with a reciprocal dependence for a given network size; furthermore, the values for the different GRNs seem to follow the functional dependence between this fraction and network size shown by Debian.

Moreover, it is possible to quantify ‘damage avalanches’ in GRNs. Gene knock-out experiments silence individual genes and monitor the cascade of genes that are affected by such a network-state change. The knock-out of a single gene is able to generate ‘avalanches’ or cascades of ‘failures’ whose sizes follow a power-law distribution with an exponent $\tau = 3/2$ (see electronic supplementary material, figure S8).

We also explored how the information content of GRNs (defined as the amount of DNA per haploid nucleus in the genome, also called C-value) correlates with the number of genes in their genome for various organisms for which information was publicly available [39]. A positive correlation between the total number of genes and the genome size has been known for a long time to hold for many living organisms [39], from prokaryotes to eukaryotes and plasmids, as well as viruses. As illustrated in figure 8, our results confirm this positive correlation. Moreover, the figure shows that viruses, plasmids and prokaryotes share the same linear trend (i.e. the genome size of the organisms is not affected by organismal complexity). For eukaryotic cells, however, information content grows faster than the number of genes (i.e. the GRN contains much more DNA per gene than expected from the number of genes). This phenomenon has been called the ‘C-paradox’ or, more recently, ‘C-value enigma’ [40], and has been explained as caused by the existence of non-coding DNA in eukaryotes [41], i.e. DNA segments that do not encode protein sequences but are believed to be important for control and regulation.

4. Discussion

Tracking the evolution of Debian software networks offers an excellent opportunity to shed light and make conjectures on how information-processing biological networks could have possibly evolved. Here, we explored the emergent architectural features of Debian networks achieved through their evolutionary dynamics. In particular, we showed that Debian networks are quite sparse, and show a scale-free structure for the out-degree distribution (representing the number of dependent packages) and a stretched-exponential in-degree distribution (representing the number of processing packages). In addition, the comparison of the releases with the ensemble of randomized networks allowed us to conclude that progressively more modular structures emerge from (and play an important role in) the development of software networks, and that this is not a simple artefact of the particular degree distribution of the networks. The increase of the modular structure of the networks for the different releases is especially significant because it suggests an effective increase of their robustness, that is, in the ability to recover from malfunctioning or damaged packages [20]. Moreover, the small average path length and its change over versions reveal that the typical distance for information transmission between any two packages remained short regardless of the growing network or their actual degree (i.e. small-world effects).

In addition, the evolutionary trend for the three hierarchical categories identified here points to an increase in the reuse of code as new software was developed, which may have resulted from seeking a more efficient package use. As a consequence, Debian is becoming progressively more ‘entangled’ [42]. The reuse of code (or software) to optimize the development of the operating system is, nonetheless, concerted with an effective and robust transfer of information, which is appreciable in the fact that the measures of hierarchy that refer to information flow show high directionality and low reciprocal dependencies. The potential for information flow improvement is compatible with the behaviour observed for l_G , as well as the power-law distribution observed for the cascade of failures originated from individual-package deletion. For the latter, an exponent $\tau = 3/2$ (approximately what we observed for the last releases) is commonly interpreted as underlying critical unbiased branching process dynamics [31,43,44], and is observed in many other contexts in biological systems [31]. However, our results stem from structural effects, without requiring critical dynamics to explain them. Moreover, while the network has become more entangled the ‘vulnerability’ of the system has decreased across releases. All together suggests that Debian has evolved towards a robust state in which code reuse is concomitant with an increase of the modularity of the network and a decrease of the risk of a collapse of the entire system [20]. Although the increase in modularity in spite of the increase in the number of pass-through packages may seem surprising, it is understandable if this reutilization of resources seems to be done ‘within moduli’ and thus preserving modular structure.

Interestingly, many of the structural properties discussed above seem to have changed their trend at a particular moment of the Debian evolutionary path. For example, the exponents of the in- and out-degree distributions, average in-degree, average path length and modularity (of the swapped versions) all reach a stationary value around version 8. This points to an evolutionary steady state (or, at least, stasis) for

the network structure. At this particular version, we also observed an acceleration in the increase of modularity of the original networks and their total size, and a deceleration in the growth of the network (number of packages) and vulnerability of the network. Finally, also at this version, we observed high peaks of variability in different quantities (e.g. Q , l_G). These peaks indicate a large variability in the ensemble of randomized networks, which reveals some type of dramatic structural reorganization (transition) of the network ensemble. We were unable to track down any specific event/change in the development history of Debian that could explain solely all this phenomenology, although some important changes were introduced in the versions around and at the eighth release. From a historical point of view, the eighth release saw dramatic design changes that were meant to make the distribution more user-friendly, and more focused by reducing the broad spectrum of architectures covered by the operating system. This release introduced the 64-bit architecture, and a more modular installation system (debian-installer) that included a tool to retrieve all dependencies for a specific package upon installation [25]. At the same time, Ubuntu (a distribution which aimed to improve the ease of the installation and use of Linux without compromising in robustness) was born, built on Debian's architecture and infrastructure. This distribution became very popular and its development led to a more professional/less hobbyist approach from developers, many of which created and maintained packages for both distributions. This feedback between Debian and Ubuntu (and change of mindset) may have also been important for the improvement in robustness that we report here. Thus, not only technical aspects but also environmental and sociological aspects have contributed to the evolution of Debian.

4.1. Analogies and differences between Debian and gene-regulatory networks

On the basis of a reductionist position, the central dogma of molecular biology states that each isolated gene—i.e. the basic unit of heredity—is transcribed into RNA. The latter, in turn, is usually translated into a protein, which can regulate the expression of other genes, usually inter-related with other proteins, thus forming a complex network of regulatory interactions [45]. This entails causal relationships and the transfer of sequential information between such genes. Therefore, genetic expression is an essential biological process that can be represented in terms of GRNs [37]. Based on causal topological relations and the transfer of sequential information, we propose that the software networks of Debian constitute a useful analogy to understand some aspects of GRNs. Many of the features described here for Debian networks are similarly found in real GRNs.

The role of software packages in Debian networks is played by genes in GRNs, while the role of software dependencies is played by genetic regulatory dependencies or interactions. Both are typically sparse networks, with a similar structure reflected in their degree distributions: a scale-free structure for the out-degree distribution and a stretched-exponential in-degree distribution, a high level of modularity, and similar (small) values for the average path length indicating small-world effects. These similarities suggest the existence of some common underlying generative mechanisms for both types of networks, maybe focused on the common goal of information transmission.

GRNs show a hierarchical structure, which has been hypothesized to confer an effective and robust way to transfer information and coordinate diverse processes [28,46], fostering the flow of information and providing effective responsiveness to external stimuli [38,47]. Differently from Debian, however, the number of pass-through information units in GRNs is very reduced, showing a more pyramidal hierarchical structure. This is a key difference between the two systems. Although an advantage when it comes to the development of new individual packages, reuse of code should *a priori* decrease its robustness, as it facilitates the spread of damage. The suggestion that the reuse of code in Debian is kept within modules allows the operating system to grow with the advantages of both the efficiency of code reuse and modularity. Thus, both systems seem to have found ways to develop robustness, Debian aiming at a cost-efficient design principle [28]. Nonetheless, both Debian and GRNs show a similar behaviour for 'damage avalanches', as the latter also show a power law distribution for avalanche size with an exponent close to $3/2$ (see above) [31,44].

Remarkably, the Debian software networks show a cross-over from linear to superlinear behaviour (figure 6) similar to that of the so-called 'C-value paradox' for GRNs (figure 8). For the latter, this change of slope is now known to result from the existence of non-coding DNA sequences. The functionality and role of such non-coding DNA, however, has not been fully elucidated and its utility or futility are still under discussion, even if it seems clear that it is important for gene regulation [48,49]. However, it is important to stress that some studies have revealed the tendency of the genomic size to increase evolutionarily because mutation sometimes duplicates parts of the genome (i.e. by means of non-adaptive processes without, in principle, any functional meaning). Some of these duplicated genes were removed by natural selection (purging non-beneficial sequences) while some others resulted in novel substrates for the evolution of further complexity and functionalities, possibly including a role in regulatory functions [50]. Interestingly, looking into the structure of individual Debian packages, we found extra pieces of information that are present in each package beyond the essential information required to work. In particular, handbooks or extra control information encapsulated in particular control files, which contain meta-information such as the needed dependencies, as well as mechanisms designed to avoid errors during unpacking and installation processes. Such information ensures a proper transmission of information during the installation process, i.e. ensures that packages are really functional in the whole network [51]. This forms for software the equivalent of non-coding pieces of DNA in GRNs. In this analogy, at least one of the roles of non-coding pieces is the monitoring and minimization of gene disruption during the transcriptional process. We hypothesize that control files and extra information comprise and ensure the resilience of the network, minimizing risks in the transmission of the information of the system. The increased growth of the amount of information in the highly complex networks of the last releases points to an increasing relevance of control files as an effective strategy in order to minimize the disruption during information processing. This could allow for an emergent optimized structure for the flow of information throughout the network. Although all these hypotheses can help understand the function and evolution of non-coding DNA, further work is needed to clarify the role and changes of control files over releases, as well as the evolution of the internal structure of software packages.

5. Conclusion

As shown here, in many aspects Debian networks are able to recreate the emergent properties observed in real GRNs [19,20,28]. In both synthetic and real biological networks, common solutions emerge to the general problem of designing circuitry that optimizes storage, information processing (of both internal and external stimuli), and robustness. Debian has evolved towards a stable scale-free out-degree distribution that is also shown by GRNs. In Debian, this property emerges from a network that grows by adding packages that are not necessarily needed by others but that require using key existing ones. For GRNs, this would mean that the genome grows by adding new genes that are controlled by key, existing genes, opening thus new regulatory pathways. In fact, empirical evidence from analysis of *E. coli* and yeast GRNs reveals that duplication of target genes has contributed far more to the growth of their GRNs than duplication of transcription factors (e.g. [52]). The exponential in-degree distribution in Debian resulted from a self-regulated limitation as to how much of the existing network is reutilized by new packages, i.e. there is a finite, well-defined mean number of dependencies per package. For GRNs, this self-regulation is convenient as it reduces the vulnerability of the network. The evolution of Debian has increased its robustness through a reduced average path length and increased modularity, which could be a key reason why current GRNs, also showing low l_G and high Q , have been selected for. The two types of networks, however, have found different strategies to increase their robustness. Whereas the growth of the Debian network has required a within-moduli reuse of packages, GRNs have a pyramidal hierarchy that could have resulted from adding new genes that introduce functional redundancy (as opposed to re-using existing ones), which considerably reduces the vulnerability of the network.

The study of similarities and differences between Debian and GRNs has helped here understand better the evolution of the latter. Conversely, it might be of interest to consider borrowing and implementing some features and strategies from real GRNs for the development of operating systems. In future work, we plan to develop a mathematical model of growing networks aimed at accounting for these common architectural features between software and GRNs, which can help identify the origin and nature of the divergences between them. Moreover, we plan to extend this study to other Linux distributions and other software ecosystems (e.g. Python Package Index or R-CRAN), with the aim to unveil key commonalities and differences that help us understand the different evolutionary solutions to common developmental challenges. Similar to the present study, such an analysis can contribute to our understanding of the evolutionary processes that current self-organized biological networks have undergone or will undergo in the future.

6. Methods

6.1. Debian networks

The Debian GNU/Linux networks are composed of software packages (units of software), which can depend on other packages (the so-called dependencies) and/or exhibit conflicts with other packages. A dependence implies a direct link (from package *A* to package *B*) that means that package *B* should be installed first to

be able to install *A* (and, consequently, information flows from *B* to *A*). For conflicts, the link means that package *A* cannot be installed if package *B* is present (i.e. installed) in the system.

Conflicts are mostly used to avoid duplicities that can ensue from the different options for the same requirement or functionality (grouped in the so-called ‘virtual packages’), as well as to select between previous versions of different packages, as explained in the Debian Policy Manual [26]. Conflicts between virtual facilities can be solved by considering particular choices among the various possibilities. Thus, in order to build Debian potential networks excluding duplication of the so-called *virtual* packages, here we selected randomly a particular choice of each possible real package providing such functionality, and obtained averages over choices (specifically, over 10^4 realizations for every distribution), hence exploring the whole ensemble of networks for each release. As a result of solving such conflicts, these Debian potential networks indicate the requirements (or dependency network) between the different packages.

6.2. Statistical analysis

We performed curve fits for the cumulative degree distributions for dependencies, employing a Levenberg–Marquardt algorithm [53]. For cases compatibles with power laws, we used cumulative distributions for this curve-fitting; in other words, if the expectation is a degree distribution of the form $P(k) \sim k^{-\alpha}$, we studied instead its cumulative version ($cP(k) = P(K \leq k) = \int_{k_{\min}}^k P(k) dk \sim cP(k) \sim K^{-\alpha+1}$). On the other hand, when the expectation was a stretched exponential (e.g. in-degree distribution), that is $P(k) \sim \exp(-(k/\tau)^\beta)$, we instead curve-fitted its transformation $\log P(k) \sim -(k/\tau)^\beta$.

In order to compare modularity and another properties across releases (see below), we also computed the Z-score with respect to random networks, with random realizations respecting the degree sequence of the network (i.e. the number of in- and out-neighbours for each package). The Z-score of an observable in the network is defined as the difference between the observable and the mean, weighted by standard deviation of the randomization: $z = (x - \mu)/\sigma$.

6.3. Network swapping

Given the original degree sequence of a Debian network (*D*), the randomization of the Debian potential networks is performed by making modifications on a copy (*D'*). Such modifications are no more than a swapping process that maintains both the in-degree and the out-degree connectivity. The swap of the links is performed as follows:

- (i) Select randomly a package *A* and one of its dependencies *a*.
- (ii) Select randomly a different package *B* and one of its dependencies *b* with the conditions: (i) *A* cannot depend on *b* and (ii) *B* cannot depend on *a*.
- (iii) Swap both links, i.e. *A* depends on *b* and *B* depends on *a*.
- (iv) Iterate this process a large number of times until the network becomes fully randomized, i.e. until the observable in the randomized network reaches a stationary state.

Once this process has been completed, the result is a randomized version of the focal Debian network/release that can be used to measure all the desired structural features, with the confidence that the networks are comparable. Here, we used an ensemble of over 10^4 realizations for every release.

6.4. Network structural features

To measure of the modularity index *Q* in the different releases of Debian we used a heuristic method, the Louvain method [54] on directed networks [55], based on modularity optimization.

The modularity index is defined as [7]

$$Q = \frac{1}{m} \sum_{i,j} \left[A_{ij} - \frac{d_i^{\text{in}} d_j^{\text{out}}}{m} \right] \delta(c_i, c_j), \quad (6.1)$$

where A_{ij} is the adjacency matrix, m is the number of edges, d_i^{in} (respect. d_i^{out}) stands for the in-degree (out-degree) of the node i and c_i is the community to which vertex i belongs. The Louvain method provided in this case an excellent accuracy due to the relatively large size of the last Debian releases.

On the other hand, the average path length on the graph (l_G , i.e. the number of steps along the shortest paths for all possible pairs of network nodes) is defined as

$$l_G = \frac{1}{m \cdot (m-1)} \cdot \sum_{i \neq j} d(n_i, n_j), \quad (6.2)$$

where m is the number of links and $d(n_i, n_j)$ is the shortest distance between nodes i and j . Here, we measured l_G using the *igraph R* package [56].

To quantify hierarchy, we used three different indicators. First, we measured the number of sinks, sources and pass-through packages, which provides an indication of hierarchical level and re-use of packages in the network. In addition, we measured an index that focuses on the directionality of the information flow in the network [29], the so-called flow index χ , which measures the fraction of links pointing from lower to higher hierarchical levels, i.e. aligned with the inherent directionality. Computing χ requires the determination *a priori* of a hierarchical ordering of all nodes in the network for which one can use the algorithm developed by Dominguez-Garcia *et al.* [29]. The index takes the value $\chi = 1$ in the limit of perfect feedforwardness (perfect directionality or hierarchical organization), and $\chi = 1/2$ in the absence of a well-

defined directionality (e.g. for a random directed network). For convenience when using a semi-logarithmic plot, we represented in our figures $1 - \chi$ instead. Lastly, we measured a version of the classical hierarchy index that quantifies the level of mutual dependencies. The Krackhardt hierarchy score K_{HS} is defined as the fraction of (unordered) node pairs (i, j) such that node i is reachable from node j but node j is *not* reachable from node i or vice versa [30]:

$$K_{\text{HS}} = 1 - \frac{V}{N(N-1)}, \quad (6.3)$$

where V is the number of symmetrically linked pairs and $N(N-1)$ is the total number of pairs in the network. Here, we instead represented $1 - K_{\text{HS}}$, which indicates the fraction of pairs that show mutual dependencies.

Data accessibility. Code and data used for the analysis of the Debian GNU/Linux and gene regulatory networks is downloadable from <https://github.com/pvgongora/Debian>.

Authors' contributions. J.A.B. and M.A.M. conceived the project. P.V. performed the numerical simulations and prepared the figures. P.V., J.A.B. and M.A.M. wrote and reviewed the manuscript.

Competing interests. We declare we have no competing interests.

Funding. We acknowledge Spanish-MINECO grant no. FIS2017-84256-P (FEDER funds) for financial support, as well as the Consejería de Conocimiento, Investigación y Universidad, Junta de Andalucía and European Regional Development Fund (ERDF), ref. SOMM17/6105/UGR.

Acknowledgements. We thank E. Estrada, P. Moretti and J. M. Martín for very useful comments. We are also grateful to two anonymous reviewers, whose comments and suggestions have greatly improved this paper.

References

- Benner SA, Sismour AM. 2005 Synthetic biology. *Nat. Rev. Gen.* **6**, 533–543. (doi:10.1038/nrg1637)
- Purnick PE, Weiss R. 2009 The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.* **10**, 410–422. (doi:10.1038/nrm2698)
- Khalil AS, Collins JJ. 2010 Synthetic biology: applications come of age. *Nat. Rev. Gen.* **11**, 367–379. (doi:10.1038/nrg2775)
- Hagmann P *et al.* 2008 Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6**, e159. (doi:10.1371/journal.pbio.0060159)
- Sole RV, Montoya M. 2001 Complexity and fragility in ecological networks. *Proc. R. Soc. Lond. B* **268**, 2039–2045. (doi:10.1098/rspb.2001.1767)
- Pastor-Satorras R, Vespignani A. 2001 Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200. (doi:10.1103/PhysRevLett.86.3200)
- Newman ME. 2003 The structure and function of complex networks. *SIAM Rev.* **45**, 167–256. (doi:10.1137/S003614450342480)
- Kitano H. 2002 Systems biology: a brief overview. *Science* **295**, 1662–1664. (doi:10.1126/science.1069492)
- Alon U. 2006 *An introduction to systems biology: design principles of biological circuits*. London, UK: CRC Press.
- Jeong H, Tombor B, Albert R, Oltvai Z, Barabasi A. 2000 The large-scale organization of metabolic networks. *Nature* **407**, 651–654. (doi:10.1038/35036627)
- Barabasi A-L, Oltvai ZN. 2004 Network biology: understanding the cell's functional organization. *Nat. Rev. Gen.* **5**, 101–113. (doi:10.1038/nrg1272)
- Kauffman S. 1969 Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467. (doi:10.1016/0022-5193(69)90015-0)
- Kauffman SA. 1993 *The origins of order: self-organization and selection in evolution*. New York, NY: Oxford University Press.
- Gros C. 2011 *Random Boolean networks*. Berlin, Germany: Springer.
- Villegas P, Ruiz-Franco J, Hidalgo J, Muñoz MA. 2016 Intrinsic noise and deviations from criticality in boolean gene-regulatory networks. *Sci. Rep.* **6**, 34743. (doi:10.1038/srep34743)
- Albert R, Othmer HG. 2003 The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J. Theor. Biol.* **223**, 1–18. (doi:10.1016/S0022-5193(03)00035-3)
- Li F, Long T, Lu Y, Ouyang Q, Tang C. 2004 The yeast cell-cycle network is robustly designed. *Proc. Natl Acad. Sci. USA* **101**, 4781–4786. (doi:10.1073/pnas.0305937101)
- Tkačik G, Callan CG, Bialek W. 2008 Information flow and optimization in transcriptional regulation. *Proc. Natl Acad. Sci. USA* **105**, 12 265–12 270. (doi:10.1073/pnas.0806077105)
- Fortuna MA, Melián CJ. 2007 Do scale-free regulatory networks allow more expression than random ones? *J. Theor. Biol.* **247**, 331–336. (doi:10.1016/j.jtbi.2007.03.017)
- Fortuna MA, Bonachela JA, Levin SA. 2011 Evolution of a modular software network. *Proc. Natl Acad. Sci. USA* **108**, 19 985–19 989. (doi:10.1073/pnas.1115960108)
- Valverde S. 2005 Logarithmic growth dynamics in software networks. *Europhys. Lett.* **72**, 858. (doi:10.1209/epl/i2005-10314-9)
- Pang TY, Maslov S. 2013 Universal distribution of component frequencies in biological and technological systems. *Proc. Natl Acad. Sci. USA* **110**, 6235–6239. (doi:10.1073/pnas.1217795110)
- Keil P *et al.* 2018 Macroecological and macroevolutionary patterns emerge in the universe of GNU/Linux operating systems. *Ecography* **41**, 1788–1800. (doi:10.1111/ecog.03424)
- Valverde S. 2015 Punctuated equilibrium in the large-scale evolution of programming languages. *J. R. Soc. Interface* **12**, 20150249. (doi:10.1098/rsif.2015.0249)
- Hertzog R, Mas R. 2014 *The Debian Administrator's Handbook, Debian Wheezy from Discovery to Mastery*.

26. Jackson I *et al.* 2017 Debian policy manual. See <http://www.debian.org/doc/manuals/debian-policy/v4.1.1.1>.
27. Corominas-Murtra B, Goñi J, Rodríguez-Caso C. 2013 On the origins of hierarchy in complex networks. *Proc. Natl Acad. Sci. USA* **110**, 13 316–13 321. (doi:10.1073/pnas.1300832110)
28. Yan K-K, Fang G, Bhardwaj N, Alexander RP, Gerstein M. 2010 Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc. Natl Acad. Sci. USA* **107**, 9186–9191. (doi:10.1073/pnas.0914771107)
29. Domínguez-García V, Pigolotti S, Muñoz MA. 2014 Inherent directionality explains the lack of feedback loops in empirical networks. *Sci. Rep.* **4**, 7497. (doi:10.1038/srep07497)
30. Krackhardt D. 1994 Graph theoretical dimensions of informal organizations. In *Computational organizational theory* (eds K Carley, M Prietula), pp. 89–111. Hillsdale, NJ: Lawrence Erlbaum Associates.
31. Muñoz MA. 2018 Colloquium: criticality and dynamical scaling in living systems. *Rev. Mod. Phys.* **90**, 031001. (doi:10.1103/RevModPhys.90.031001)
32. Gama-Castro S *et al.* 2015 Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* **44**, D133–D143. (doi:10.1093/nar/gkv1156)
33. Albergante L, Blow JJ, Newman TJ. 2014 Buffered qualitative stability explains the robustness and evolvability of transcriptional networks. *Elife* **3**, e02863. (doi:10.7554/eLife.02863)
34. Arrieta-Ortiz ML *et al.* 2015 An experimentally supported model of the *Bacillus subtilis* global transcriptional regulatory network. *Mol. Syst. Biol.* **11**, 839. (doi:10.15252/msb.20156236)
35. Ma S, Kemmeren P, Gresham D, Statnikov A. 2014 De-novo learning of genome-scale regulatory networks in *S. cerevisiae*. *PLoS ONE* **9**, e106479. (doi:10.1371/journal.pone.0106479)
36. Aldana M, Balleza E, Kauffman S, Resendiz O. 2007 Robustness and evolvability in genetic regulatory networks. *J. Theor. Biol.* **245**, 433–448. (doi:10.1016/j.jtbi.2006.10.027)
37. Buchanan M. 2010 *Networks in cell biology*. Cambridge, UK: Cambridge University Press.
38. Ravasz E, Somera AL, Mongru DA, Oltvai ZN. 2002 Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555. (doi:10.1126/science.1073374)
39. Geer LY *et al.* 2009 The NCBI biosystems database. *Nucleic Acids Res.* **38**, D492–D496. (doi:10.1093/nar/gkp858)
40. Gregory TR. 2001 Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev.* **76**, 65–101. (doi:10.1017/S1464793100005595)
41. Ohno S. 1972 So much 'junk' DNA in our genome. In *Brookhaven Symp. Biol.* **23**, 366–370.
42. Donetti L, Neri F, Muñoz MA. 2006 Optimal network topologies: expanders, cages, Ramanujan graphs, entangled networks and all that. *J. Stat. Mech. Theory Exp.* **2006**, P08007. (doi:10.1088/1742-5468/2006/08/P08007)
43. di Santo S, Villegas P, Burioni R, Muñoz MA. 2017 Simple unified view of branching process statistics: random walks in balanced logarithmic potentials. *Phys. Rev. E* **95**, 032115. (doi:10.1103/PhysRevE.95.032115)
44. Rämö P, Kesseli J, Yli-Harja O. 2006 Perturbation avalanches and criticality in gene regulatory networks. *J. Theor. Biol.* **242**, 164–170. (doi:10.1016/j.jtbi.2006.02.011)
45. Crick F. 1970 Central dogma of molecular biology. *Nature* **227**, 561–563. (doi:10.1038/227561a0)
46. Wagner GP, Pavlicev M, Cheverud JM. 2007 The road to modularity. *Nat. Rev. Gen.* **8**, 921–931. (doi:10.1038/nrg2267)
47. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. 2004 Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **14**, 283–291. (doi:10.1016/j.sbi.2004.05.004)
48. Carey N. 2015 *Junk DNA: a journey through the dark matter of the genome*. New York, NY: Columbia University Press.
49. Mercer TR, Dinger ME, Mattick JS. 2009 Long non-coding RNAs: insights into functions. *Nat. Rev. Gen.* **10**, 155–159. (doi:10.1038/nrg2521)
50. Lynch M, Conery JS. 2003 The origins of genome complexity. *Science* **302**, 1401–1404. (doi:10.1126/science.1089370)
51. Krafft MF. 2005 *The Debian system: concepts and techniques*. San Francisco, CA: No Starch Press.
52. Teichmann SA, Babu MM. 2004 Gene regulatory network growth by duplication. *Nat. Genet.* **36**, 492–496. (doi:10.1038/ng1340)
53. Moré JJ. 1978 The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis* (ed. GA Watson). Lecture Notes in Mathematics, vol. 630, pp. 105–116. Berlin, Germany: Springer.
54. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008 Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008. (doi:10.1088/1742-5468/2008/10/P10008)
55. Dugué N, Perez A. 2015 Directed Louvain: maximizing modularity in directed networks. PhD thesis, Université d'Orléans.
56. Csardi G, Nepusz T. 2006 The igraph software package for complex network research. *InterJournal Complex Systems* **CX.18:1695**, 1–9.