Percolation in bipartite Boolean networks and its role in sustaining life

R Hannam, R Kühn, and A Annibale

Department of Mathematics, King's College London,
The Strand, London WC2R 2LS, United Kingdom

Abstract. Boolean networks are popular models for gene regulation, where genes are regarded as binary units, that can be either expressed or not, each updated at regular time intervals according to a random Boolean function of its neighbouring genes. Stable gene expression profiles, corresponding to cell types, are regarded as attractors of the network dynamics. However, the random character of gene updates gives no insight into the biological mechanism behind the existence of attractors. We propose a bipartite Boolean network approach which integrates genes and regulatory proteins (i.e. transcription factors) into a single network, where interactions incorporate two fundamental aspects of cellular biology, i.e. gene expression and gene regulation, and the resulting dynamics is highly non-linear. Since any finite stochastic system is ergodic, the emergence of an attractor structure, stable under noisy conditions, requires a giant component in the bipartite graph. By adapting graph percolation techniques to directed bipartite graphs, we are able to calculate exactly the region, in the network parameters space, where a cell can sustain steady-state gene expression profiles, in the absence of inhibitors, and we quantify numerically the effect of inhibitors. Results show that for cells to sustain a steady-state gene expression profile, transcription factors should typically be small protein complexes that regulate many genes. This condition is crucial for cell reprogramming and remarkably well in line with biological facts.

E-mail: alessia.annibale@kcl.ac.uk

1. Introduction

Originally introduced by Kauffman to model gene regulatory networks in living cells [1], Boolean networks have since become one of the most popular class of models to analyze complex systems of interacting units, finding use in a wide variety of fields including spin-glasses [2], neural networks [3], computing circuits [4], time-series [5], biological [6, 7], economic [8] and geological sciences [9]. Their dynamic properties, in particular their phase transition behaviour and number and size of attractors, have been studied extensively [10, 12, 11, 13, 14, 15] and have been shown to display features of biological systems, such as evolvability [16], homeostasis [17] and criticality [18], and to capture much of the phenomenology of single gene knock-out experiments [19]. However, the random character of the update functions of its constituent units, does not carry any information on the biological mechanism with which genes interact with each other, making it difficult to establish an explicit link between gene interactions and the existence of cellular attractors, let alone transitions between them, as observed in cellular differentiation or in cell reprogramming experiments [20], although general mechanisms for cell differentiation have been proposed [34, 31]. In conventional

Boolean networks, each gene is updated according to a randomly chosen Boolean function of the states of neighbouring genes, which are themselves picked at random. Cell types are then rationalised as attractors of a dynamical system of many interacting degrees of freedom, however, no hint is given on how and why genes interact.

Recently, neural network models of gene regulation, encoding cellular attractor structures in the interactions between regulatory genes (or proteins), have been successful in modelling dynamics similar to those observed in experiments [21, 22, 23]. However, the dense nature of the (Hopfield-like) interactions, assumed in these models, is at odds with biological facts. Indeed, each gene is known to be regulated by a small number of transcription factors (TFs), that are single (or small complexes of) proteins, and each protein is synthetised from one gene. This leads to effective gene-gene interactions, mediated by TFs, which are sparse and directed.

Although gene-regulatory networks (GRNs) and protein interaction networks have traditionally been studied separately, they are deeply connected: gene interactions exist only *through* and *via* TFs. This induces a special structure in GRNs, in particular a precise relation must exist between the degree statistics of GRNs and the statistics of TF sizes (i.e. number of proteins that consitute a TF). In this work, we propose a bipartite Boolean model which integrates regulatory genes and TFs into a single bipartite network, with *sparse* and *directed* links, encoding two fundamental aspects of cell biology, i.e. gene expression (of proteins forming TFs) and regulation of genes (by TFs). Recently, bipartite networks with sparse interactions have been shown to have learning and parallel retrieval capabilities, via their link with restricted Boltzmann machines [24, 25] and neural networks with diluted patterns [26, 35]. However, this connection only holds for bipartite networks with *symmetric* interactions. Processing capabilities of sparse directed bipartite networks are still unexplored territory. Our analysis will show that for a non-trivial gene expression profile to be stable in noisy conditions, characteristic of biological systems, TFs need to be typically single proteins or small protein complexes, able to regulate many genes. This condition is a *prerequisite* for the existence of cellular life, and it is remarkably well in line with biological findings.

The rest of this paper is organised as follows: in Sec.2 we will give an account of the intricate dynamics of gene expression, protein synthesis, TF formation and gene regulation; in Sec.3 we will introduce a bipartite graph formalism that simplifies the description; in Sec.4 we will solve the model and in Sec.5 we will discuss results and pathways for future work. Technical details are relegated to appendices.

2. Gene expression, protein synthesis, complex formation and gene regulation

Gene expression is regulated by transcription factors (TFs) that are proteins or complexes of proteins, which are themselves synthetised from expressed genes. One can visualise the network of regulatory genes and TFs as a multilayered network (see Figure 1) with layers representing the different components involved in gene regulation: genes, proteins and protein complexes. Some interactions, like gene expression (from genes to proteins) and gene regulation (from proteins or protein complexes to genes), are directed, whilst others, e.g. proteins forming a complex, are non-directed. We describe the gene expression level of each gene i by a binary variable σ_i taking the value 1 if the gene is expressed, and 0 if it is not. Gene expression levels are updated at regular time intervals, measured e.g. in terms of stages of the cell cycle. We can

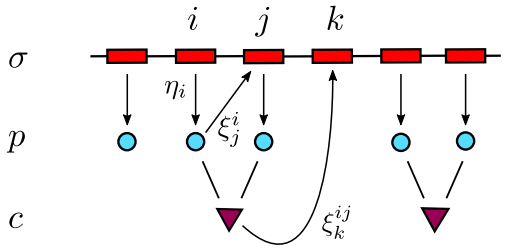

Figure 1. A network representation of gene regulation (not all nodes/edges are shown). A gene σ_i synthesises a protein p_i, which can reversibly bind (undirected edges) to form protein complexes c_{ij}. The proteins and complexes may act as transcription factors (TFs) of the genes. The regulatory effect that a TF has on a gene is given by $\xi_j^{a(b)} \in \{0, \pm 1\}$: the superscript of ξ denotes the gene(s) contributing to the TF synthesis, whilst the subscript denotes the gene regulated by the TF.

express its dynamics in terms of the concentration of proteins p_j, binary complexes c_{jk} and higher order complexes that regulate its expression, through the multilayered network parameters, as:

$$\sigma_i(t+1) = \Theta\left[\sum_j \xi_i^j p_j + \sum_{j,k} c_{jk}\xi_i^{jk} + \ldots - Tz_i - \theta_i\right] \tag{1}$$

Here Θ is the Heaviside step function, z_i is a zero-averaged random variable mimicking biological noise, T is a parameter that scales the strength of the noise, $\xi_i^a \in \{0, \pm 1\}$ denotes the regulatory effect of TF a (protein or complex) on gene i, and θ_i is a gene specific threshold, representing a barrier that regulatory interactions need to overcome to activate the gene. The ellipsis represent regulatory effects from higher order complexes that we do not write explicitly. Reaction equations, in continuous time, can be written for the evolution of the concentration of proteins,

$$\dot{p_i} = \sigma_i\eta_i - p_i\sum_j p_j\Pi_{ij}^+ + \sum_j c_{ij}\Pi_{ij}^- - \gamma_i p_i\,, \tag{2}$$

and protein complexes,

$$\dot{c_{ij}} = p_i p_j\Pi_{ij}^+ - c_{ij}\Pi_{ij}^- - \gamma_{ij}c_{ij}\,. \tag{3}$$

Here η_i is the rate of protein i synthesis, $\Pi_{ij}^{\pm}$ are association/dissociation rates for the complex of protein i and j and the $\boldsymbol{\gamma}$ variables are degradation rates. The protein-protein interaction network of a cell can be constructed from the non-zero values of Π_{ij}^+ that details which proteins interact with one another. Assuming that the dynamics of protein synthesis, dissociation and decay is much faster than that of gene expression, one can apply a separation of time scales resulting in stationarity in $\mathbf{p}$ and $\mathbf{c}$ for each time step in the gene expression dynamics. At stationarity (3) gives,

$$c_{ij} = \frac{\Pi_{ij}^+}{\Pi_{ij}^- + \gamma_{ij}} p_i p_j\,, \tag{4}$$

which allows us to write the stationary solution of (2) as

$$p_i = \frac{\sigma_i\eta_i}{\sum_j \frac{\gamma_{ij}\Pi_{ij}^+}{\Pi_{ij}^- + \gamma_{ij}} p_j + \gamma_i}\,. \tag{5}$$

When $\sigma_i = 0$, the protein from gene i is not synthesised and (5) rightly gives $p_i = 0$. When $\sigma_i = 1$ expanding the right hand side of (5) in the limit of small protein concentration (i.e. $p_j = 0$) gives,

$$p_i \simeq \frac{\eta_i}{\gamma_i}\left(1 - \frac{1}{\gamma_i}\sum_j \frac{\gamma_{ij}\Pi^+_{ij}}{\Pi^-_{ij}+\gamma_{ij}} pj\right), \tag{6}$$

which can be rearranged into the following form,

$$\gamma_i \simeq \sum_j \left[\frac{\gamma_i^2}{\eta_i}\delta_{ij} + \frac{\gamma_{ij}\Pi^+_{ij}}{\Pi^-_{ij}+\gamma_{ij}}\right] p_j\,, \tag{7}$$

or, in vector notation,

$$\boldsymbol{\gamma} = \mathbf{M}\mathbf{p}, \tag{8}$$

where $\mathbf{M}$ is a matrix with symmetric and diagonal parts, $\mathbf{M} = \mathbf{S} + \mathbf{D}$. By inverting this matrix equation, stationary values of protein concentrations are found as

$$p_i = \sigma_i \sum_j M^{-1}_{ij}\gamma_j\,. \tag{9}$$

The expansion for small protein concentrations carried out above relies on the fact that concentrations of transcription factors are usually quite low (typically in the nM range) and comparable to their dissociation factors [27].

Substituting into (1), one obtains the following equations for gene expression levels which are effectively interacting through J_{ij}, J_{ijk} and higher order interactions

$$\sigma_i(t{+}1) = \Theta\left[\underbrace{\sum_j \xi_i^j \sum_k M^{-1}_{jk}\gamma_k}_{J_{ij}}\sigma_j + \underbrace{\sum_{j,k}\xi_i^{jk}\sum_{\ell,\ell'}\frac{\Pi^+_{jk}\gamma_\ell\gamma_{\ell'}M^{-1}_{j\ell}M^{-1}_{k\ell'}}{\Pi^-_{jk}+\gamma_{jk}}}_{J_{ijk}}\sigma_j\sigma_k + \ldots - Tz_i - \theta_i\right]$$

These interactions are sparse, as the regulatory interactions $\boldsymbol{\xi}$ are sparse. From the form of J_{ij} and J_{ijk}, it can be seen that the interaction between two genes require regulation via single proteins and the interaction between three genes requires regulation through a complex of two proteins. Expanding this reasoning, one sees that effective interactions between n genes requires complexes that are formed from $n-1$ proteins.

In the next section we will introduce a more economical description, that will allow one to go beyond binary complexes by explicitly taking into account higher order complexes. This will be achieved by regarding proteins and protein complexes, of any order, that regulate gene expressions, simply as TFs, which are formed from the expression of genes. Hence, the intricate dynamics of gene expression, protein synthesis, complex formation and gene regulation will be reduced to a bipartite graph model, with two sets of nodes, representing, respectively, genes and TFs.

3. Definition of a simplified model

3.1. Bipartite graph formulation

We model the combined network of genes and TFs as a bipartite graph, where one set of nodes is constituted by N regulatory genes and the other by $P = \alpha N$ TFs,

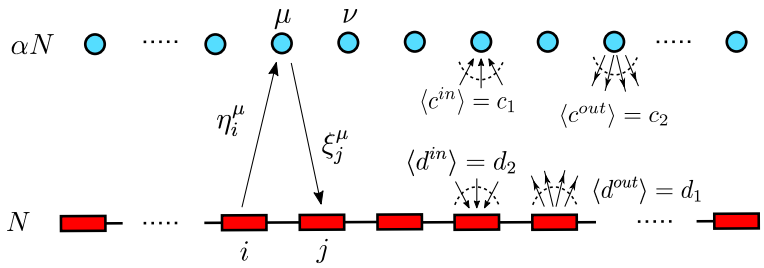

Figure 2. Bipartite graph representation of a system with N regulatory genes and αN TFs and two sets of directed links, i.e. $\boldsymbol{\eta}$ (from genes to TFs) and $\boldsymbol{\xi}$ (from TFs to genes). The average in-degrees of TFs and genes are c_1 and d_2 respectively. The average out-degrees follow from conservation of links, as $d_1 = \alpha c_1$ and $c_2 = d_2/\alpha$.

with $\alpha = \mathcal{O}(1)$‡. Each gene is modelled as a Boolean variable σ_i, where $i = 1, \ldots, N$, taking value 1 if gene i is expressed and 0 otherwise, and each TF is modelled by a real variable τ_μ, $\mu = 1, \ldots, \alpha N$, which denotes the concentration of TF μ.

A directed link $\eta_i^\mu = 1$ exists, from gene i to TF μ, if gene i expresses a protein that contributes to TF μ and a directed link $\xi_i^\mu = \pm 1$ exists, from TF μ to gene i, if TF μ activates $(+1)$ or inhibits (-1) the expression of gene i (see Fig.2). One has $\eta_i^\mu = 0$, if no link exists from gene i to TF μ, and $\xi_i^\mu = 0$, if no link exists from TF μ to gene i. This formulation is similar to the one introduced in gene protein Boolean networks (GPBN) [32, 33], however, there is a key difference: a gene does not necessarily produce just a single TF, but may contribute to several different ones through the formation of complexes that contain the protein encoded by the gene in question.

For a given directed bipartite network $(\boldsymbol{\xi}, \boldsymbol{\eta})$, we denote the in-degree of TF μ by $c_\mu^{\rm in}(\boldsymbol{\eta}) = |\partial_\mu^\eta|$, where $\partial_\mu^\eta = \{i : \eta_i^\mu = 1\}$, and the in-degree of gene i by $d_i^{\rm in}(\boldsymbol{\xi}) = |\partial_i^\xi|$, with $\partial_i^\xi = \{\mu : |\xi_i^\mu| = 1\}$. For simplicity, we consider random graph ensembles with the nonzero ξ's and η's independently and identically distributed according to $P(\eta = 1) = c_1/N$ and $P(|\xi| = 1) = d_2/\alpha N$, with $c_1 = \mathcal{O}(1)$ and $d_2 = \mathcal{O}(1)$. Then, the resulting distributions of the in-degrees $P_d^{\rm in}(d)$ and $P_c^{\rm in}(c)$ are Poissonian with average d_2 and c_1 respectively. Conservation of links demands $d_1 = \alpha c_1$ and $d_2 = \alpha c_2$, where d_1 and c_2 are the average out-degree of genes and TFs respectively. The out-degree distributions $P_d^{\rm out}(d), P_c^{\rm out}(c)$ are Poissonian with average d_1 and c_2, respectively. More general graph ensembles are considered in Appendix A.

Physically, c_1 represents the average size of TFs (i.e. average number of proteins forming a TF), while c_2 represents the average number of genes regulated by a TF (or TF promiscuity). It is possible to reduce the bipartite network down to a gene-gene interaction network (normally the focus of experimental work), by simply integrating out the TFs. Since TFs are the intermediaries of gene-gene interactions, there exist a precise relation between the (in- and out-) degree distribution of the gene-gene interaction network and the distribution of TF sizes and promiscuities. This relation is derived in Appendix A for general graph ensembles.

‡ There are about $N = 2500$ regulatory genes in the human genome and the number of TFs is estimated to be of the same order [28, 29, 30].

3.2. Regulatory dynamics

We assume that gene expression levels are updated synchronously via

$$\sigma_i(t+1) = \Theta\left[\sum_\mu b_i^\mu \xi_i^\mu \tau_\mu(t) - \theta_i - T z_i\right], \tag{10}$$

where b_i^μ is the binding affinity of TF μ to its target gene i. For a TF μ to be synthesised, all the genes coding for proteins that form TF μ, must be expressed at the same time. Upon introducing the order parameter

$$m_\mu(t) = \sum_j \frac{\eta_j^\mu \sigma_j(t)}{c_\mu} \tag{11}$$

which takes value 1 when this condition is satisfied, the evolution of TF concentrations can be modelled by the set of differential equations

$$\dot{\tau}_\mu = \Pi_\mu^+ \delta_{m_\mu(t),1} - \Pi_\mu^- \tau_\mu \tag{12}$$

where the production/degradation rates $\Pi_\mu^\pm$ will be set to 1 for all μ henceforth, and $\delta_{x,y}$ is the Kronecker delta. A separation of timescales, arising from the fast evolution of TFs when compared to gene expression, allows to approximate TF concentrations with their stationary values,

$$\tau_\mu(t) = \delta_{m_\mu(t),1} \tag{13}$$

Within this approximation, each τ_μ is modelled as a Boolean variable, which takes value 1 if TF μ is synthetised and 0 otherwise. From now on we will set all the binding affinities $b_i^\mu = 1\ \forall\ i$, $T = 0$ and $\theta_i = 0\ \forall\ i$, so to reduce the number of parameters explored. Substituting these values and (13) into (10), leads to the highly non-linear gene expression dynamics

$$\sigma_i(t+1) = \Theta\left[\sum_\mu \xi_i^\mu \delta_{m_\mu(t),1}\right]. \tag{14}$$

Rewriting $\delta_{m_\mu(t),1} = \prod_{j\in\partial_\mu^\eta} \sigma_j(t)$, reveals that TFs that are products of $n-1$ expressed genes create effective n-body interactions in the gene expression dynamics. In fact, one can write (14) in terms of many-body interactions $J^{(\ell)}_{ij_1\ldots j_\ell} = \sum_\mu \xi_i^\mu \eta_{j_1}^\mu \ldots \eta_{j_\ell}^\mu / c_\mu^\ell$

$$\sigma_i(t+1) = \Theta\left[\sum_{\ell\geq 1} \frac{1}{\ell!} \sum_{j_1,\ldots,j_\ell} J^{(\ell)}_{ij_1\ldots j_\ell} \sigma_{j_1}(t)\ldots\sigma_{j_\ell}(t)\right]. \tag{15}$$

Interestingly, truncation of the sum to $\ell = 1$, gives a dynamics with two-body interactions $J^{(1)}_{ij} = \sum_\mu \xi_i^\mu \eta_j^\mu / c_\mu$ of the Hebbian type. Were $\boldsymbol{\eta}=\boldsymbol{\xi}$, this would describe a neural network capable of retrieving all of its (diluted) patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^{\alpha N}$ in parallel, for $\alpha c_1^2 < 1$ [35, 36]. However, here the (diluted) Hebbian interactions are *asymmetric* and yield an interesting dynamics in its own right, that we shall refer to as "linear". The linear dynamics corresponds to replacing $\tau_\mu(t) = \delta_{m(t),1}$ in (10) with $\tau_\mu(t) = m_\mu(t)$. This will lead to a bound on the full dynamics (14), that we shall refer to as "non-linear". Note that in the linear case τ_μ is a positive real variable that measures the fraction of expressed genes among those needed to syntethise TF μ, as opposed to the non-linear dynamics where τ_μ is Boolean. It will be convenient to define a variable $\bar{\tau}_\mu(t) = \Theta[\tau_\mu(t)]$§ that indicates, for both types of dynamics, when a TF is synthetised.

§ The convention $\Theta(0) = 0$ is used.

In the early stages of the development of an embryo, the maternally inherited TFs must kick-start the zygotic gene expression dynamics [37]. A key question is under which conditions will the introduction of a small number of TFs into an inactive GRN, result in a non-trivial steady-state gene expression profile, allowing a cell to express all the necessary proteins to carry out its function and sustain its life.

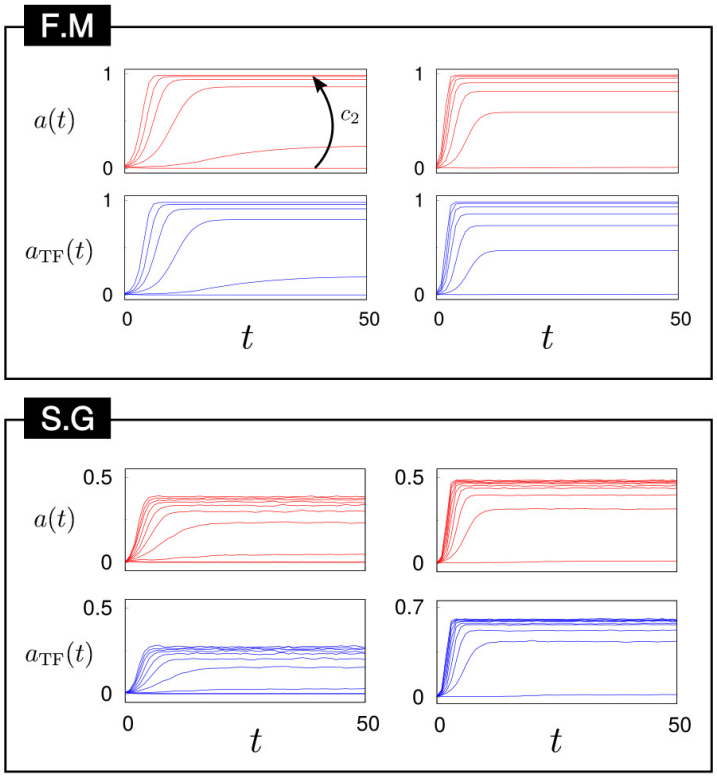

Figure 3. Average gene and TF activities, $a(t)$ and $a_{\rm TF}(t)$ in simulations of bipartite regulatory networks with $N = 2,500$, $\alpha = 1$, $c_1 = 1$, evolving via non-linear (left column) and linear dynamics (right column), from initial states where all genes are inactive and 10 TFs are present. In the top panel all of the TFs are activators (FM interactions), while in the bottom panel TFs are activators and inhibitors with equal probability (SG interactions). Each curve shows the average over 100 networks with the same connectivity c_2, ranging from $c_2 = 1, 2, \ldots 7$ (FM) and $c_2 = 1, 2, \ldots, 20$ (SG), respectively.

4. Model analysis and results

Simulations of the linear and non-linear dynamics from initial states where all genes are inactive and a few randomly selected TFs are present, show that the density of expressed genes $a(t) = N^{-1} \sum_i \sigma_i(t)$ and of synthetised TFs $a_{\rm TF}(t) = (\alpha N)^{-1} \sum_\mu \bar{\tau}_\mu(t)$,

reach a steady state that depends on the interplay between activators ($\xi = 1$) and inhibitors ($\xi = -1$) and on the network degrees (Fig.3). When all TFs are activators (i.e. interactions are ferromagnetic (FM)) and the connectivity c_2 is large, the system approaches a steady state where approximately all the genes are expressed. However, if inhibitory TFs exist (i.e. interactions are spin-glass (SG)), the dynamics plateaus to a lower value of the activity, with fluctuations occurring due to competing TFs. Furthermore, for both FM and SG interactions, the system remains in a silent state, $a(t) = \mathcal{O}(N^{-1})$, for values of the connectivity below a certain threshold.

4.1. Percolation thresholds

A crucial property of any cell is its ability to maintain a non-trivial gene expression profile. To be able to do so under noisy conditions, this must be supported by a *large* connected network of genes and TFs: noisy dynamics on any finite network is ergodic and thus unable to sustain non-trivial expression profiles [38, 12]. This implies that the relevant regulatory network must reside on the giant cluster (GC) of the bipartite network. In addition, reprogramming experiments have recently shown that almost any cell can be reprogrammed from a *small* set of TFs (Yamanaka factors). To induce macroscopic changes in cellular gene expression profiles, these must *necessarily* hit the GC. Hence, the question of percolation (i.e. emergence of a GC) becomes of paramount importance. Percolation is known to describe the phase transition of a ferromagnet on a finitely connected network in the absence of noise and has also been studied for directed networks [39, 40] and undirected bipartite graphs [39, 41]. Here, we calculate exactly the percolation threshold for a directed bipartite network, where all TFs are activators, i.e. $\xi_i^\mu \in \{0, 1\}$, and genes are updated via the *non-linear* Boolean function (14), using an adaptation of the cavity method [42]. To this end, we introduce indicator variables n_i, n_μ, that take value 1 if gene i or TF μ belong to the GC, respectively, and are zero otherwise. The indicator variable for any gene or TF can be expressed in terms of the corresponding indicator variables for their neighbouring nodes in the cavity graph,

$$n_i = 1 - \prod_{\mu \in \partial_i^\xi} \left(1 - n_\mu^{(i)}\right), \tag{16}$$

$$n_\mu = \prod_{j \in \partial_\mu^\eta} n_j^{(\mu)}. \tag{17}$$

Here ∂_i^ξ represents the nodes that are the nearest neighbours of gene i via a ξ-edge, and $n_\mu^{(i)}$ is the indicator variable for TF μ in the cavity graph, where gene i and all the edges connecting to it are removed. Similarly, ∂_μ^η denotes the nearest neighbours of TF μ connected to one of its η-edges, and $n_j^{(\mu)}$ is the indicator variable for gene j on the cavity graph with TF μ removed. Note that in (17) and (19) below, it is required that empty products are evaluated to zero instead of one. Equation (17) reflects the *non-linear character* of the dynamics: transcription factor μ belongs to the GC if and only if all the genes contributing to its synthesis are on the GC. In contrast, from (16), a gene i belongs to the GC if at least one of its regulating TFs belongs to the GC. Similarly, the cavity equations for the nearest-neighbours of i and μ read

$$n_i^{(\mu)} = 1 - \prod_{\nu \in \partial_{i\setminus\mu}^\xi} \left(1 - n_\nu^{(i)}\right), \tag{18}$$

$$n_\mu^{(i)} = \prod_{k \in \partial^\eta_{\mu \setminus i}} n_k^{(\mu)} \tag{19}$$

where $\partial^\xi_{i \setminus \mu}$ denotes the set of nodes connected to i via ξ-edges, excluding the node μ (similarly for $\partial^\eta_{\mu \setminus i}$). Equations (18), (19) are exact on trees and, in the thermodynamic limit, they will also become exact on graphs sampled from our ensemble, which are locally tree-like due to the sparsity of ξ_i^μ and η_i^μ. In this limit, we can average (16) and (17) over the graph ensemble. By using the generating functions $G^{(d,c)}(x) = \sum_{k=1}^\infty P^{\text{in}}_{d,c}(k) x^k$, and assuming that $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are independent, we obtain for the probability $g = \langle n_i \rangle$ and $t = \langle n_\mu \rangle$ of being in the GC

$$1 - g = G^{(d)}\left(1 - \tilde{t}\right), \qquad t = G^{(c)}(\tilde{g}) \tag{20}$$

where the cavity probabilities $\tilde{g} = \langle n_i^{(\mu)} \rangle$, $\tilde{t} = \langle n_\mu^{(i)} \rangle$ solve the self-consistency equations (B.3, B.4) and can be shown to be identical, up to differences $\mathcal{O}(N^{-1})$, to g, t, respectively, due to the sparsity, directedness and independence of the links (see Appendix B).

A stability analysis shows that a non-zero solution, corresponding to the emergence of a GC, exists when the average TF out-degree is above a critical value, given, for Poisson degree distributions, by (see Appendix B)

$$c_2^* = \frac{e^{c_1}}{\alpha c_1}. \tag{21}$$

The percolation threshold for the linear dynamics is found in a similar way in Appendix B. Here equations (17) and (19) take the same form as (16) and (18), respectively, and lead to

$$c_2^* = \frac{1}{\alpha c_1}. \tag{22}$$

This generalizes existing results for the percolation threshold in symmetric bipartite graphs [35, 36], that are retrieved by setting $c_2 = c_1$. We note that (22) could also have been obtained by marginalising the bipartite network over the TFs to obtain an effective gene-gene interaction network with average degree $\alpha c_1 c_2$ (as demonstrated in Appendix A), and then applying known results about percolation in directed graphs [40]. Indeed, in the absence of inhibitors, the linear dynamics describes the evolution of a ferromagnet on the marginalised gene-gene interaction network. We note, however, that this simplification does not arise for the non-linear dynamics.

Plots of the percolation threshold c_2^* for the linear and non-linear dynamics are shown in fig.4 (left panel). Notably, the region where a GC exists, is much wider for the linear than for the non-linear dynamics. This is as expected, since to activate a (TF) node in the non-linear dynamics it is required that all of the neighbouring (gene) nodes are active. This describes a bootstrap process on directed bipartite graphs. *Bootstrap percolation* [43] has been studied extensively on lattices [44, 45, 46, 47], regular graphs [48, 49], trees [50], and, recently, on complex networks [51, 52], however there are no exact results on bipartite graphs. The uncovered asymmetry in the roles of c_1 and c_2, is remarkably well in line with biological facts [53]. TFs are typically formed by a small number $c_1 \in \{1, 2, 3\}$ of proteins and tend to regulate a large number $c_2 \in \{10, \ldots, 100\}$ of genes, while physiological values of α are around one.

Phase diagrams for non-Poissonian degree distributions can be derived directly from equations (B.3), (B.4) for the non-linear and (B.13), (B.14) for the linear dynamics, performing the stability analysis in Appendix B, whilst degree correlations

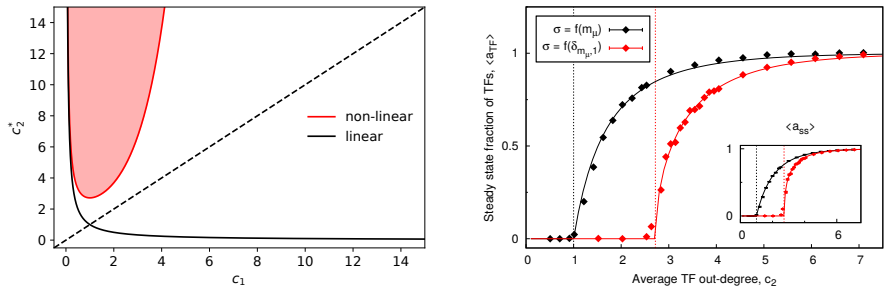

Figure 4. Left: Average TF out-degree c_2^* at which a giant cluster (GC) emerges in the directed bipartite network as a function of the average TF in-degree c_1 (for $\alpha = 1$). The non-linear dynamics allows a GC in the red-shaded region above the (dashed) line $c_2 = c_1$, which is a subset of the region where a GC exists for the linear dynamics (above the solid black line). Right: Average steady state fraction of TFs versus their average out-degree c_2, in bipartite networks with Poisson degree distributions and $c_1 = 1$, evolving via nonlinear (black line) and linear (red line) dynamics, respectively. Solid lines show the analytical solution (B.10) of equations (20) for the non-linear dynamics, and the solution (B.16) of the corresponding equations for the linear dynamics. Symbols show simulation results for $N = 2500$. The vertical dashed lines indicate the theoretically predicted percolation thresholds. The inset shows the average steady state gene expression levels for the same simulations.

can be accounted for in (B.1), (B.2) and (B.11), (B.12), respectively. Non-Poissonian degree distributions and correlations in the links are expected to impact the phase diagram, leading to a more complex dependence of the GC on the degree statistics.

Simulation results for the average steady-state fractions of expressed genes $\langle a_{ss} \rangle$ and synthetised TFs $\langle a_{\rm TF} \rangle$ are shown in fig.4 (right panel), for both linear and non-linear dynamics, for bipartite graphs with $N = 2500$ and Poisson degree statistics, initialised in a state where all genes are off and a small number ($\sim$10) of TFs, randomly selected, are present. Phase transitions occur at the theoretically predicted values of c_2^*. Below the percolation threshold, the network is disconnected, thus introducing a small set of TFs will only activate a small number of genes, while above the percolation threshold, this will result in an activation "avalanche", due to the presence of a GC. Simulations are in excellent agreement with the theoretically predicted values of g and t, given, for Poisson distributions, by the analytical expressions (B.10) for the non-linear dynamics, and (B.16) for the linear dynamics. It is worth noting that, because the networks generated for simulations were constructed using Poisson degree distributions, there are more nodes simulated than those that can be interpreted as regulatory genes or TFs. This is because a regulatory gene must contribute to the synthesis of at least one TF, i.e. $d_i^{\rm out} \geq 1$. In addition, every gene (including regulatory genes) must be regulated by at least 1 TF, i.e. $d_i^{\rm in} \geq 1$. Similarly, any TF must be synthesised by at least one regulatory gene, i.e $c_\mu^{\rm in} \geq 1$, and then can regulate the expression of any gene either in the regulatory part of the network or outside of this sub-network, so $c_\mu^{\rm out} \geq 0$. In fig. 4, both the analytic solutions and the simulation results take this into account and are normalised by the appropriate probabilities. For example, $\langle a_{\rm TF} \rangle$ is normalised by $P(c^{\rm in} > 0)$. All of these simulations were carried out deterministically ($T = 0$) and with no gene-specific thresholds $\theta_i = 0 \ \forall i$. In addition, up until this point, only *activating* TFs have been included in the dynamics.

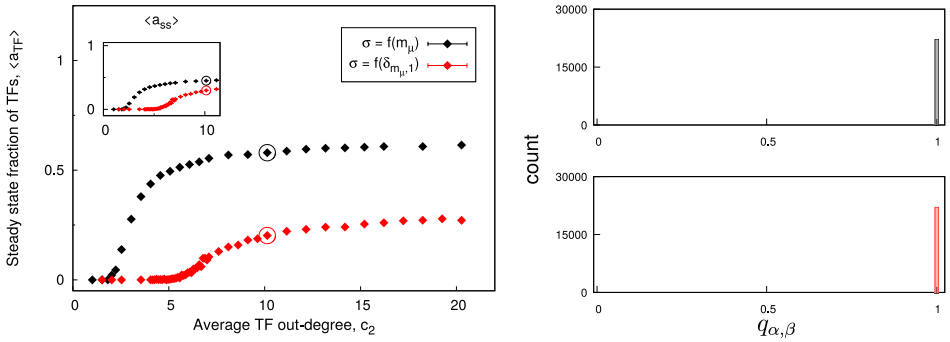

Figure 5. Simulation results for the linear (black) and non-linear (red) dynamics of bipartite networks with $N = 2500$, $\alpha = 1$, $c_1 = 1$, $\epsilon = 0.5$ and Poisson degree distribution. Left: fraction of TFs in the steady-state. The fraction of active genes is shown in the inset. Right: distribution of the overlap $q_{\alpha\beta}$ over 150 simulation runs of the linear (top panel) and non-linear (bottom panel) dynamics for the c_2 value highlighted on the left panel. Self-overlaps $q_{\alpha\alpha}$ are not plotted.

The effect of inhibitors will be analysed in the next subsection, where TFs will be assumed to activate or inhibit their target genes with probability $P(\xi = 1) = \epsilon$ and $P(\xi = -1) = 1 - \epsilon$ respectively.

4.2. The effects of inhibition

In the above derivations, we have assumed that all TFs are *activators*. The presence of inhibitors is expected, on general grounds, to shrink the stability region, in the parameters space, of cellular attractors, and to lower the steady state fractions of expressed genes and synthetised TFs. This is consistent with simulation results shown in fig.5 (left panel) for regulatory interactions with nonzero $\xi_i^\mu = \pm 1$ equally likely for all i, μ, i.e. $\epsilon = 0.5$.

In addition, competition between excitatory and inhibitory interactions is expected to introduce frustration in the system, as TFs compete to regulate the same genes in different manners. However, due to the asymmetry of interactions, no multiplicity of attractors is expected [54, 55, 56, 57]. This is confirmed by computing the distribution of overlaps $q_{\alpha\beta}$ between the steady-state TF trajectories attained in different simulations of the same network, for different initial conditions (i.e. sets of 4 randomly selected TFs). The overlap, measuring the similarity of the steady-state TF profiles between two simulation runs, α and β, is defined as $q_{\alpha\beta} = \tilde{q}_{\alpha\beta}/\sqrt{\tilde{q}_{\alpha\alpha}\tilde{q}_{bb}}$, where

$$\tilde{q}_{\alpha\beta} = \frac{1}{P}\sum_{\mu}\left(\langle\bar{\tau}_\mu\rangle_\alpha - a_\alpha,\right)\left(\langle\bar{\tau}_\mu\rangle_\beta - a_\beta\right) , \tag{23}$$

$\langle\ldots\rangle_\alpha$ denotes the time average in the steady state of simulation run α and $\bar{\tau}_\mu$ indicates whether a TF has been synthesised or not. In the above,

$$a_\alpha = \frac{1}{P}\sum_{\mu}\langle\bar{\tau}_\mu\rangle_\alpha , \tag{24}$$

so that the overlap is defined as a Pearson correlation coefficient, taking values in $[-1, 1]$. When $\beta = \alpha$, $q_{\alpha\alpha} = 1$ and the TF profiles are identical. Any value of $q_{\alpha\beta} \neq 1$

indicates that there is a difference in the steady-state TF profiles between different simulation runs. The overlap distribution $P(q_{\alpha\beta})$ is shown in the right panels of fig.5 for both the linear (top panel) and the non-linear (bottom panel) dynamics. This was computed for $M = 150$ different simulations for an arbitrary point in the parameter space above c_2^*, ensuring that the network has a non-zero $\langle a_{\rm TF}\rangle$, for both versions of the dynamics. The plot does not show the self-overlaps ($q_{\alpha\alpha} = 1$), in order to focus on overlaps between different simulation runs. For both the linear and non-linear dynamics the distribution of overlaps has a single peak at $q_{\alpha\beta} = 1$. This implies that despite the SG characters of interactions, in both types of dynamics, each network supports a single attractor.

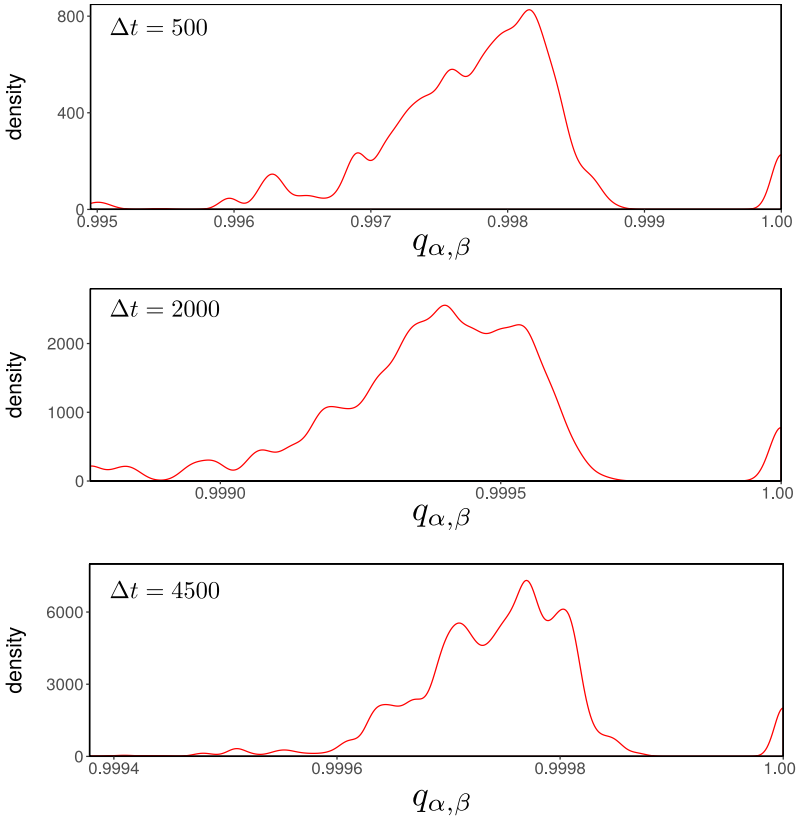

Figure 6. The probability density function of the overlap between the same 150 simulations of the non-linear deterministic gene regulatory dynamics on a fixed network with $c_1 = 1$, $c_2 = 10$, $\epsilon = 0.5$. The different distributions arise from different time windows over which the steady state dynamics was averaged: $\Delta t = 500$ (top), $\Delta t = 2,000$ (middle), $\Delta t = 4,500$ (bottom).

Looking closer at the structure of the probability density function (pdf) of the overlap, near $q_{\alpha\beta} = 1$, shows that the distribution is not a perfect δ-function at exactly $q_{\alpha\beta} = 1$. However, increasing the time window over which the steady state average is performed moves the mass of the pdf towards $q_{\alpha\beta} = 1$, suggesting the existence of a single limit cycle attractor that either has a long period or a short period that is traversed many times (figure 6). The period can be inferred from the

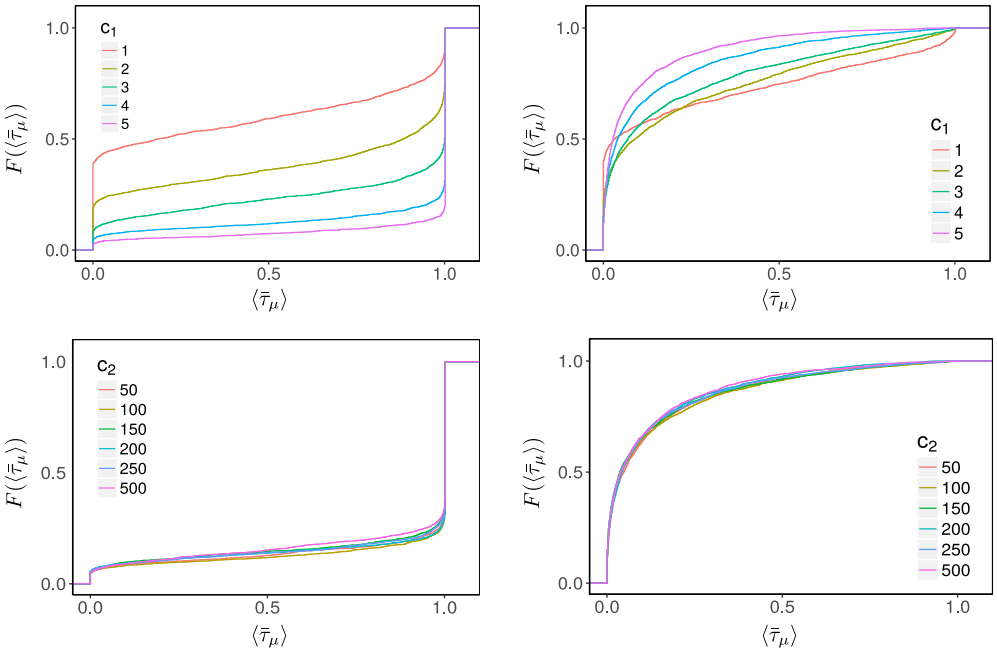

Figure 7. Empirical cumulative distribution functions (CDFs) of the steady state frequency that a TF is synthesised in bipartite networks with different connectivities and $\epsilon = 0.5$, for non-linear (left panels) and linear (right panels) dynamics. Top panels: $c_2 = 100$. Bottom panels: $c_1 = 4$.

average deviations of the overlap from $q_{\alpha\beta} = 1$, as shown in Appendix C. Results suggest the existence of a single limit cycle with a short period. This is supported by a recent study [58] where the effect of dilution and asymmetry of interactions on the mean number and mean length of limit cycles attractors was explored numerically in neural network models analogous to the linear dynamics considered in this work. Similarly to what is observed in random Boolean networks and diluted neural networks [12, 59], the study suggests that increasing the asymmetry of interactions dramatically decreases the number of limit cycle attractors, whilst increasing the sparsity of the interactions decreases the length of the limit cycles. It is also shown that there is a dramatic decrease in the mean number of attractors as the interactions become very sparse. However, at high levels of dilution, this may be due to the lack of a GC in the networks.

The impact of the connectivities c_1, c_2 and of the bias ϵ on the trajectories is shown in Fig. 7 and 8, respectively, where empirical cumulative distribution functions (CDF) of the steady state frequency of TF synthesis, $\langle\bar{\tau}_\mu\rangle = \frac{1}{\Delta t}\sum_{t=t'}^{t'+\Delta t}\bar{\tau}_\mu(t)$ are plotted for the linear (left panels) and non-linear dynamics (right panels). In the non-linear dynamics, increasing c_1 decreases the likelihood that each TF will be synthesised, for any choice of ϵ, while in the linear dynamics the opposite is true, i.e. increasing c_1 increases the probability that a TF is expressed in the steady state (top panels of Fig. 7). On the other hand, increasing c_2 has a remarkably small effect in both dynamics (bottom panels). Finally, increasing ϵ increases the frequency with which each gene is expressed, and thus each TF is synthesised in the steady state (Fig. 8).

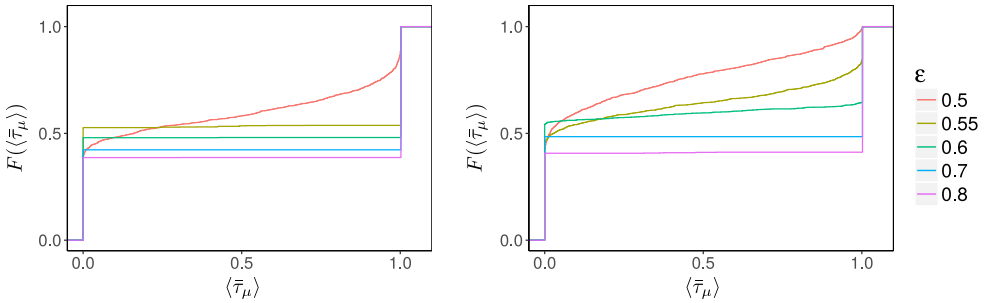

Figure 8. Empirical CDF of the steady state frequency with which TFs are synthesised, using the linear (left) and non-linear (right) dynamics on a network with $(c_1, c_2) = (1, 10)$ at $T = 0$. As the bias towards activation increases the number of TFs which are always expressed increases.

Because any given network with a giant cluster was observed to support a single gene expression profile in the steady state, this may suggest that different cell types may require different networks, e.g. different rates of protein production, degradation and TF binding affinities. However, we note that a multiplicity of states can also be likely achieved by the same network, with a different prescription for the regulatory interactions ξ_i^μ (that here have been set to $0, \pm 1$). Indeed the current choice is only a subset of the possible regulatory interactions that can arise in random Boolean networks, where a multiplicity of states is observed. We intend to investigate this in greater detail in future work.

5. Summary and Outlook

GRNs and transcription networks have been traditionally studied separately, although they are deeply related. Models of GRNs, including Boolean networks and neural networks, have been successful in capturing different aspects of gene expression dynamics, however they provide little insight into the biological mechanism underpinning the existence of cellular attractors. In this work, we have proposed a bipartite Boolean modelling approach to gene regulation, which integrates regulatory genes and TFs into a single bipartite network, with sparse and directed links, encoding two fundamental aspects of cell biology, i.e. gene expression (of proteins forming TFs) and regulation of genes (by TFs). The resulting dynamics is highly non-linear and it describes a Bootstrap process, where several genes have to be simultaneously expressed to synthetise a TF. In order for such non-linear dynamics to sustain a non-trivial gene expression profile under noisy conditions characteristic of biological systems, the combined network of genes and TFs *needs to have* a giant component. This requires TFs to be typically small protein complexes that regulate many genes. This condition is remarkably well in line with biological findings and it is at the root of reprogramming experiments, where a small set of TFs is observed to drastically change the gene expression profile of a cell.

There are several pathways for future work. Firstly, only deterministic dynamics has been studied in this work, so stochasticity effects should be considered in future studies. These are expected to restrict the range, in the parameter space, where a "frozen" phase with non-trivial gene expression profiles emerges, however, stochasticity

may also enrich the range of behaviours supported in the frozen phase. Furthermore, the assumption that $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are statistically independent was made. This may not be true, in particular c^{in} and c^{out} might be correlated, as the number of DNA binding sites may increase with the size of a TF, hence the inclusion of correlations may make the model more realistic. Also, the model does not take into account the effects of external signals (e.g. morphogen gradients and cell-cell interactions), which could be included, in future developments, via additional external fields. In our view, the most fruitful advancement of this model would be the introduction of suitable weights for the regulatory interactions, that are able to embed a multiplicity of attractors, as it is required for multi-cellular life. Another possibility is to investigate this model on temporal networks where edges evolves in response to variations in the rates for protein synthesis/degradation and TF binding affinities. This may also create a multiplicity of attractors. In addition, if such values were dependent on gene expression profiles it may be possible for the dynamics to traverse from one attractor to another, encapsulating changes in cell state, for example, due to differentiation.

Finally, from a more theoretical point of view, the giant component considered in this work corresponds to the out-component of the strongly connected component, traditionally defined in directed graphs [39]. Existing studies [39, 40] suggest that the condition for the emergence of a giant out-component is identical to the one for the in-component. However, this equivalence is only expected to hold for the linear dynamics. Thus, it would be an interesting pathway for future work to compare these conditions for the non-linear dynamics, where the concept of in-component generalizes less easily.

Acknowledgments

RH is supported by the EPSRC Centre for Doctoral Training in Cross-Disciplinary Approaches to Non-Equilibrium Systems (CANES EP/L015854/1). All authors thank Giuseppe Torrisi for useful discussions.

References

[1] S.A. Kauffman (1969), *J. Theor. Biol.* 22(3): 437-467.
[2] P.W. Anderson (1983), *Proc. Natl. Acad. Sci. USA* 80, 33863390.
[3] J.J. Hopfield (1982) *Proc. Natl Acad. Sci. USA* 79, 25542558. D. J. Amit, H. Gutfreund, H. Sompolinsky (1985) *Phys. Rev. Lett.* 55, 1530.
[4] A. Mozeika, D. Saad, and J. Raymond (2009), *Phys. Rev. Lett.* 103, 248701.
[5] N. Berestovsky, L. Nakhleh (2013). *PLoS ONE* 8:e66031.
[6] D. Thieffry, AM Huerta, E Pérez-Rueda, J Collado-Vides (1998), *Bioessays.* 20(5):433-40.
[7] R. Thomas (1973), *J. Theor. Biol.* 42(3):563-585.
[8] J.-W. Gu, T.-K. Siu, and H. Zheng (2013), *Risk and Decision Analysis* 4(2): 119129.
[9] I. Zaliapin, V. Keilis-Borok, and M. Ghil (2003) J. Stat. Phys. 111(3-4): 839861.
[10] R. Albert, A.L. Barabási (2000). *Phys. Rev. Lett.* 84, 5660-5663.
[11] B. Derrida, Y. Pomeau (1986) *Euro. Phys. Lett.* 1, 4549. B. Derrida, D. Stauffer (1986), *Euro. Phys. Lett.* 2, 739745.
[12] U. Bastolla, G. Parisi (1996), *Physica D* 98, 125. U. Bastolla, G. Parisi (1997), *J. Theor. Biol.* 187, 117133. U. Bastolla, G. Parisi (1998), *Physica D* 115, 203218. U. Bastolla, G. Parisi (1998), *Physica D* 115, 219233.
[13] B. Luque, R.V. Solé (1997) *Phys. Rev. E* 55, 257-260.
[14] B. Drossel, T. Mihaljev, F. Greil (2005), *Phys. Rev. Lett.* 94, 088701.
[15] K. Iguchi, S. Kinoshita, H.S. Yamada (2007), *J. Theor. Biol.* 247: 138151
[16] T. Yu, J. Miller (2001), Neutrality and the Evolvability of Boolean Function Landscape. Ed. J.

Miller *et al.* Genetic Programming. EuroGP 2001. Lecture Notes in Computer Science, 2038. Springer, Berlin, Heidelberg.
[17] A. Samal and S. Jain, *BMC Systems Biology* 20082:21.
[18] R. Serra, M. Villani, A. Graudenzi, S.A. Kauffman (2007) *J. Theor. Bio.* **246**(3):449-460.
[19] R.Serra, M.Villani A.Semeria (2004) *J. Theor. Bio.* **227**(1):149-157
[20] K. Takahashi and S. Yamanaka (2006), *Cell.* 126(4): 66376.
[21] A.H. Lang, H. Li, J.J. Collins, and P. Mehta (2014), *PLoS Comput. Biol.* 10(8):1-13.
[22] R. Hannam, A. Annibale, and R. Kühn (2017), *J. Phys. A Math. Theor.* 50(42).
[23] A. Szedlak *et al.* (2017), *PLOS Comput. Biol.* 13(11):e1005849.
[24] A. Barra, G. Genovese, P. Sollich, D. Tantari (2018), *Phys. Rev. E.* 97, 022310.
[25] J. Tubiana and R. Monasson (2017) *Phys. Rev. Lett.* 118, 138301.
[26] E. Agliari, *et al.* (2012). *Phys. Rev. Lett.* 109, 268101.
[27] R Milo, P Jorgensen, U Moran, G Weber, M. Springer *BioNumbers*, the database of key numbers in molecular and cell biology. Nucleic Acids Res. 2010, 38(Database issue):D750D753.
[28] ES Lander, *et al.* (2001), *Nature* 409, 860921.
[29] JC Venter, *et al.* (2001) *Science* 291(5507):13041351.
[30] JM Vaquerizas, SK Kummerfeld, SA Teichmann, NM Luscombe (2009) *Nat. Rev. Genet.* 10, 252263.
[31] A. Wagner (1994) *Proc. Natl. Acad. Sci.* 91(10):4387-91. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization.
[32] R. Graudenzi, *et al.* (2011) *J. Comput. Bio.* **18**(10)
[33] Graudenzi, R. Serra, M. Villani, A. Colacci, S.A. Kauffman (2011) *J. Comput. Bio.* **18**(4)
[34] M. Villani, A. Barbieri, R. Serra (2011) PLoS One. 6(3): e17703
[35] P. Sollich, D. Tantari, A. Annibale, A. Barra (2014) *Phys. Rev. Lett. 113 (23), 238106.*
[36] E. Agliari, A. Annibale, A. Barra, A.C.C. Coolen, D. Tantari, 2013. *J. Phys. A: Math. Theor:* 46 (41).
[37] W. Tadros, H.D. Lipshitz (2009). *Development* 136(18):3033-42.
[38] S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes Rev. Mod. Phys. 80, 1275
[39] M.E.J. Newman, S.H. Strogatz, D.J. Watts (2001) *Phys.Rev. E*, **64** 026118
[40] M. Boguna and M.A. Serrano (2005). *Phys. Rev. E* 72(1):016106.
[41] H. Hooyberghs, B. Van Schaeybroeck, and J.O. Indekeu (2010) *Phys. A: Stat. Mech. its Appl.* 389(15):2920-2029.
[42] M. Mezard, G. Parisi (2001), *Eur. Phys. J. B* **20**, 217.
[43] J. Chalupa, P.L. Leath and G.R. Reich (1979). *J. Phys. C: Solid State Physics*, 12(1): L31.
[44] M. Aizenman and J.L. Lebowitz (1988), *J. Phys. A* 21(19): 38013813.
[45] R.H. Schonmann (1992), *Ann. Probab.* 20(1): 174-193.
[46] A.E. Holroyd (2003), *Probab. Theory Relat. Fields* 125, 195.
[47] J. Balogh and B. Bollobas (2006) *Probab. Theory Relat. Fields* 134, 624.
[48] J. Balogh and B.G. Pittel (2007), *Random Structures Algorithms* 30, 257.
[49] L.R.G. Fontes and R.H. Schonmann (2008), J. Stat. Phys. 132, 839.
[50] J. Balogh, Y. Peres, and G. Pete (2006), *Combin. Probab. and Comput.* 15, 715.
[51] G.J. Baxter, S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes (2010), *Phys. Rev. E* 82, 011103.
[52] J. Gao, T. Zhou and Y. Hu (2015), *Scientific Reports* 5: 14662.
[53] V. Narang, *et al.* (2015), *PLoS Comput. Biol.* 11(9):e1004504.
[54] H. Rieger, M. Schreckenberg, J. Zittartz (1989)*Z. Phys. B - Condensed Matter* **74**, 527-538.
[55] U. Bastolla, G. Parisi (1997) *J. Phys. A: Math. Gen.* **30** 5613
[56] Yu-qiang Ma, Chang-de Gong (1992) *Phys. Rev. B* **46**, 3436
[57] H. Sompolinsky, A. Crisanti, and H.J. Sommers (1988) *Phys. Rev. Lett.* **61** 259
[58] V. Folli, G. Gosti, M. Leonetti, G. Ruocco (2018) *Neural Networks* **104**:50-59
[59] S Hwang, *et al.*, J. Stat. Mech. (2019) 053402
[60] T. Zhou, J. Ren, M. Medo, and Y.C. Zhang (2007) *Phys. Rev. E* **76**, 046115
[61] A. Annibale, A. C. C. Coolen, N. Planell-Morell (2015) *J. R. Soc. Interface* **12**: 20150573.
[62] D.V. Filho and D.R.J. O'Neale (2018) *Phys. Rev. E* **98**, 022307

Appendix A. General bipartite graph ensembles

We have restricted ourselves to Poisson graph ensembles for simplicity, however our analysis can be easily extended to graph ensembles with arbitrary in- and out-degree

sequences

$$p(\boldsymbol{\eta}) = \prod_{i,\mu}\left[\frac{c_\mu^{\rm in}d_i^{\rm out}}{N\langle d^{\rm out}\rangle}\delta_{\eta_i^\mu,1}+\left(1-\frac{c_\mu^{\rm in}d_i^{\rm out}}{N\langle d^{\rm out}\rangle}\right)\delta_{\eta_i^\mu,0}\right]$$
$$p(|\boldsymbol{\xi}|) = \prod_{i,\mu}\left[\frac{c_\mu^{\rm out}d_i^{\rm in}}{N\langle d^{\rm in}\rangle}\delta_{|\xi_i^\mu|,1}+\left(1-\frac{c_\mu^{\rm out}d_i^{\rm in}}{N\langle d^{\rm in}\rangle}\right)\delta_{|\xi_i^\mu|,0}\right]. \quad \text{(A.1)}$$

Here $\delta_{x,y}$ is the Kronecker delta and factorisation over i and μ follows from the assumption that the edges in the network are independent. In large graphs drawn from ensemble (A.1), each gene in-degree $d_i^{\rm in}(\boldsymbol{\xi})$ is a Poissonian variable with average $d_i^{\rm in}$ and each TF in-degree $c_\mu^{\rm in}(\boldsymbol{\eta})$ is a Poissonian variable with average $c_\mu^{\rm in}$ (similarly for the out-degrees). In what follows, we will assume that the degree sequences are drawn from arbitrary distributions $p_d(d^{\rm in},d^{\rm out}) = N^{-1}\sum_i \delta_{d^{\rm in},d_i^{\rm in}}\delta_{d^{\rm out},d_i^{\rm out}}$ and $p_c(c^{\rm in},c^{\rm out}) = P^{-1}\sum_\mu \delta_{c^{\rm in},c_\mu^{\rm in}}\delta_{c^{\rm out},c_\mu^{\rm out}}$, with averages $d_1 = \langle d^{\rm out}\rangle$, $d_2 = \langle d^{\rm in}\rangle$, $c_1 = \langle c^{\rm in}\rangle$, $c_2 = \langle c^{\rm out}\rangle$. The Poissonian case considered in the main text corresponds to the choice $c_\mu^{\rm in} = c_1, c_\mu^{\rm out} = c_2\ \forall\ \mu$ and $d_i^{\rm in} = d_2, d_i^{\rm out} = d_1\ \forall\ i$. For the general ensemble (A.1), equations (B.3), (B.4), (B.13), (B.14) remain valid, but apply to the distributions $P_d^{\rm in}(d) = N^{-1}\sum_i \pi_{d_i^{\rm in}}(d)$ and $P_c^{\rm in}(c) = P^{-1}\sum_\mu \pi_{c_\mu^{\rm in}}(c)$, where $\pi_c(k) = e^{-c}c^k/k!$ is the Poissonian distribution with average c.

Appendix A.1. Degree distribution for gene-gene interaction networks

Transcription factors act as intermediaries in a gene-regulatory network. An effective gene to gene interaction network can be obtained as $A_{ij} = \Theta\left[\sum_\mu \eta_i^\mu|\xi_j^\mu|\right]$, i.e. by integrating out the TFs (as done similarly in [39, 60, 61, 62]). Here, a directed edge exists from i to j ($A_{ij}=1$) if gene i expresses a protein forming a TF that regulates gene j. The distribution of the in- and out-degrees in the "projected" gene-gene interaction network can be calculated as follows. For the out-degrees we have:

$$p_{\rm out}(k) = \left\langle \frac{1}{N}\sum_i \delta_{k,\sum_j A_{ij}}\right\rangle_{\boldsymbol{\eta},\boldsymbol{\xi}} \quad \text{(A.2)}$$

where the average is taken over (A.1). Using the Fourier represenation of δ-functions

$$p_{\rm out}(k) = \frac{1}{N}\sum_i \int\frac{d\omega}{2\pi}e^{{\rm i}\omega k}\langle e^{-{\rm i}\omega\sum_j A_{ij}}\rangle_{\boldsymbol{\eta},\boldsymbol{\xi}}. \quad \text{(A.3)}$$

One can simplify the calculation by replacing the binary link $A_{ij} = \Theta[\sum_\mu \eta_i^\mu|\xi_i^\mu|]$ with the weighted link $\tilde{A}_{ij} = \sum_\mu \eta_i^\mu|\xi_i^\mu|$. This approximation is exact to $\mathcal{O}(N^{-1})$ as the probability that $\tilde{A}_{ij} > 1$ is $\mathcal{O}(N^{-2})$, as shown below:

$$\begin{aligned}
p(\tilde{A}_{ij}) = \langle\delta_{\tilde{A}_{ij},\sum_\mu \eta_i^\mu|\xi_j^\mu|}\rangle &= \int\frac{d\omega}{2\pi}e^{{\rm i}\omega\tilde{A}_{ij}}\langle e^{-{\rm i}\omega\sum_\mu \eta_i^\mu|\xi_j^\mu|}\rangle_{\boldsymbol{\eta},\boldsymbol{\xi}}\\
&= \int\frac{d\omega}{2\pi}e^{{\rm i}\omega\tilde{A}_{ij}}\prod_\mu\langle\eta_i^\mu|\xi_j^\mu|e^{-{\rm i}\omega}+1-\eta_i^\mu|\xi_j^\mu|\rangle_{\boldsymbol{\eta},\boldsymbol{\xi}}\\
&= \int\frac{d\omega}{2\pi}e^{{\rm i}\omega\tilde{A}_{ij}}\prod_\mu\left[\frac{d_i^{\rm out}c_\mu^{\rm in}}{N\langle d^{\rm out}\rangle}\frac{d_j^{\rm in}c_\mu^{\rm out}}{N\langle d^{\rm in}\rangle}(e^{-{\rm i}\omega}-1)+1\right]\\
&= \int\frac{d\omega}{2\pi}e^{{\rm i}\omega\tilde{A}_{ij}}e^{\sum_\mu\frac{d_i^{\rm out}c_\mu^{\rm in}}{N^2\langle d^{\rm out}\rangle}\frac{d_j^{\rm in}c_\mu^{\rm out}}{\langle d^{\rm in}\rangle}(e^{-{\rm i}\omega}-1)+\mathcal{O}(N^{-2})},
\end{aligned}$$

where the sparse nature of the graph was used to exponentiate in the last line. Expanding the exponential and using the definition of the δ-function

$$p(\tilde{A}_{ij}) = \delta_{\tilde{A}_{ij},0}\left[1 - \sum_\mu \frac{d_i^{\rm out} d_j^{\rm in} c_\mu^{\rm out} c_\mu^{\rm in}}{N^2 \langle d^{\rm out}\rangle \langle d^{\rm in}\rangle}\right] + \delta_{\tilde{A}_{ij},1}\left[\sum_\mu \frac{d_i^{\rm out} d_j^{\rm in} c_\mu^{\rm out} c_\mu^{\rm in}}{N^2 \langle d^{\rm out}\rangle \langle d^{\rm in}\rangle}\right] + \mathcal{O}(N^{-2})$$

shows that $p(\tilde{A}_{ij} > 1) = \mathcal{O}(N^{-2})$, hence to order N^{-1}, $p(\tilde{A}_{ij}) = p(A_{ij})$, and averages of A_{ij} can be replaced with averages of $\tilde{A}_{ij}$. Thus,

$$\begin{aligned}
\langle e^{-{\rm i}\omega \sum_j A_{ij}}\rangle_{\boldsymbol{\eta},\boldsymbol{\xi}} &= \prod_\mu \langle e^{-{\rm i}\omega \eta_i^\mu \sum_j |\xi_j^\mu|}\rangle_{\boldsymbol{\eta},\boldsymbol{\xi}} \qquad \text{(A.4)}\\
&= \prod_\mu \left[1 + \frac{d_i^{\rm out} c_\mu^{\rm in}}{N\langle d^{\rm out}\rangle}\left(\langle e^{-{\rm i}\omega \sum_j |\xi_j^\mu|}\rangle - 1\right)\right]\\
&= \prod_\mu \left[1 + \frac{d_i^{\rm out} c_\mu^{\rm in}}{N\langle d^{\rm out}\rangle}\left(\prod_j \left[1 + \frac{d_j^{\rm in} c_\mu^{\rm out}}{N\langle d^{\rm in}\rangle}\left(e^{-{\rm i}\omega} - 1\right)\right] - 1\right)\right]\\
&= \prod_\mu \left[1 + \frac{d_i^{\rm out} c_\mu^{\rm in}}{N\langle d^{\rm out}\rangle}\left(\exp\left\{\frac{1}{N}\sum_j \frac{d_j^{\rm in} c_\mu^{\rm out}}{\langle d^{\rm in}\rangle}\left(e^{-{\rm i}\omega} - 1\right)\right\} - 1\right)\right]\\
&= \prod_\mu \left[1 + \frac{d_i^{\rm out} c_\mu^{\rm in}}{N\langle d^{\rm out}\rangle}\left(\exp\{c_\mu^{\rm out}(e^{-{\rm i}\omega} - 1)\} - 1\right)\right]\\
&= \exp\left\{\frac{1}{N}\sum_\mu \frac{d_i^{\rm out} c_\mu^{\rm in}}{\langle d^{\rm out}\rangle}\left(\exp\{c_\mu^{\rm out}(e^{-{\rm i}\omega} - 1)\} - 1\right)\right\}\\
&= e^{\frac{-d_i^{\rm out}\alpha\langle c^{\rm in}\rangle}{\langle d^{\rm out}\rangle}} e^{\frac{\alpha d_i^{\rm out}}{\langle d^{\rm out}\rangle}\langle c^{\rm in} e^{c^{\rm out}(\exp(-{\rm i}\omega)-1)}\rangle},
\end{aligned}$$

where, the average $\langle c^{\rm in} e^{c^{\rm out}(\exp(-{\rm i}\omega)-1)}\rangle$ is taken over the joint distribution $p(c^{\rm in}, c^{\rm out})$. Substituting this result into our expression for $p_{\rm out}(k)$ gives

$$p_{\rm out}(k) = \frac{1}{N}\sum_i e^{-d_i^{\rm out}} \int \frac{d\omega}{2\pi} e^{{\rm i}\omega k} \exp\left[\frac{\alpha d_i^{\rm out}}{\langle d^{\rm out}\rangle}\langle c^{\rm in} e^{c^{\rm out}(e^{-{\rm i}\omega}-1)}\rangle\right]$$

Upon introducing the marginal distribution $p_d^{\rm out}(d^{\rm out}) = \sum_{d^{\rm in}} p(d^{\rm in}, d^{\rm out})$ and assuming independence of TF in- and out-degrees, $p(c^{\rm in}, c^{\rm out}) = p_c^{\rm in}(c) p_c^{\rm out}(c)$, one gets, using conservation of links $\langle d^{\rm out}\rangle = \alpha\langle c^{\rm in}\rangle$

$$\begin{aligned}
p_{\rm out}(k) &= \sum_d e^{-d} p_d^{\rm out}(d) \int \frac{d\omega}{2\pi} e^{{\rm i}\omega {\rm k}} \exp\left[d\left\langle \exp\left(c^{\rm out}(e^{-{\rm i}\omega} - 1)\right)\right\rangle\right]\\
&= \sum_\lambda \frac{1}{\lambda!}\sum_d d^\lambda e^{-d} p_d^{\rm out}(d) \int \frac{d\omega}{2\pi} e^{{\rm i}\omega {\rm k}} \left[\sum_c p_c^{\rm out}(c) \exp\left(c(e^{-{\rm i}\omega} - 1)\right)\right]^\lambda\\
&= \sum_\lambda \frac{1}{\lambda!}\sum_d p_d^{\rm out}(d) d^\lambda e^{-d} \sum_{c_1 \ldots c_\lambda} p_c^{\rm out}(c_1)\ldots p_c^{\rm out}(c_\lambda) \int dx \delta\left(x - \sum_{r=1}^{\lambda} c_r\right) e^{-x}\frac{x^k}{k!}
\end{aligned}$$

Thus, the out-degree distribution in the effective gene-gene interaction network is related to the distribution of TF promiscuities $p_c^{\rm out}(c)$ via

$$p_{\rm out}(k) = \int dx e^{-x}\frac{x^k}{k!} P(x)\,,$$

with

$$P(x) = \sum_d p_{\rm out}(d) e^{-d} \sum_{\lambda \geq 0} \frac{d^\lambda}{\lambda!} \sum_{c_1 \ldots c_\lambda} p_{\rm out}(c_1) \ldots p_{\rm out}(c_\lambda) \delta \left(x - \sum_{r=1}^{\lambda} c_r \right) .$$

Clearly, the out-degree distribution is normalised $\sum_{k\geq 0} p_{\rm out}(k) = 1$ and its average is

$$\begin{aligned} \langle k_{\rm out} \rangle &= \sum_{k \geq 0} k p_{\rm out}(k) = \int_0^\infty dx P(x) e^{-x} x \sum_{k \geq 0} \frac{x^{k-1}}{(k-1)!} \\ &= \int_0^\infty x P(x) dx = \sum_d p_{\rm out}(d) e^{-d} \sum_{\lambda \geq 0} \frac{d^\lambda}{\lambda!} \sum_{c_1 \ldots c_\lambda} p_{\rm out}(c_1) \ldots p_{\rm out}(c_\lambda) \sum_{r \leq \lambda} c_r \\ &= \sum_d p_{\rm out}(d) e^{-d} \sum_{\lambda \geq 0} \frac{d^\lambda}{\lambda!} \lambda \sum_c p_{\rm out}(c) c = \langle c^{\rm out} \rangle \langle d^{\rm out} \rangle = \alpha c_1 c_2 . \end{aligned} \tag{A.5}$$

It can be shown in a similar fashion that the in-degree distribution for the effective gene-gene interaction network is related to the distribution of TF sizes $p_c^{\rm in}(c)$ via

$$p_{\rm in}(k) = \int dy e^{-y} \frac{y^k}{k!} P(y) ,$$

with

$$P(y) = \sum_d p_{\rm in}(d) e^{-d} \sum_{\lambda \geq 0} \frac{d^\lambda}{\lambda!} \sum_{c_1 \ldots c_\lambda} p_{\rm in}(c_1) \ldots p_{\rm in}(c_\lambda) \delta \left(y - \sum_{r=1}^{\lambda} c_r \right) .$$

As before, one can easily show that $\langle k^{\rm in} \rangle = \alpha c_1 c_2 \equiv \langle k^{\rm out} \rangle$, as it should. For the Poissonian ensemble considered in the main text, where $p_c^{\rm in}(c) = \delta_{c,c_1}$, $p_c^{\rm out}(c) = \delta_{c,c_2}$, $p_d^{\rm in}(d) = \delta_{d,d_2}$ and $p_d^{\rm out}(d) = \delta_{d,d_1}$, the above expressions simplify to

$$p_{\rm out}(k) = e^{-d_1} \sum_\ell \frac{d_1^\ell}{\ell!} e^{-\ell c_2} \frac{(\ell c_2)^k}{k!} = \sum_\ell \pi_{d_1}(\ell) \pi_{\ell c_2}(k) \tag{A.6}$$

$$p_{\rm in}(k) \;\; = e^{-d_2} \sum_\ell \frac{d_2^\ell}{\ell!} e^{-\ell c_1} \frac{(\ell c_1)^k}{k!} = \sum_\ell \pi_{d_2}(\ell) \pi_{\ell c_1}(k) \tag{A.7}$$

where $\pi_c(k)$ is the Poissonian distribution with average c.

Appendix B. Percolation thresholds

Here the critical value c_2^* of the TF out-degree, above which a giant cluster (GC) will exist in the bipartite network, is calculated for both the nonlinear and the linear dynamics.

Appendix B.1. Non-linear dynamics

In the construction of the cavity graph for $n_\mu^{(i)}$, as given in (19), one removes the gene i connected to TF μ via an η-edge. Given that, from (16), gene i is connected to TF μ via ξ_i^μ, the likelihood that gene i is also connected to TF μ via an η-edge is $\mathcal{O}(N^{-1})$, due to the sparsity of links. In particular, for TF μ to belong to the GC in the cavity graph, where its *successor* i has been removed, all of the c_μ^{in} *predecessors* of μ must be

in the GC. Averaging (19) over all possible TFs that have i as a successor (i.e. have a link that terminates at node i) leads to

$$\tilde{t} = \sum_{c^{\rm out}} P_c^{\rm out}(c^{\rm out}) \frac{c^{\rm out}}{\langle c^{\rm out}\rangle} \sum_{c^{\rm in}=1}^{\infty} P(c^{\rm in}|c^{\rm out}) \tilde{g}^{c^{\rm in}}, \tag{B.1}$$

where the average is taken over the likelihood $P_c^{\rm out}(c^{\rm out})c^{\rm out}/\langle c^{\rm out}\rangle$ to pick up a predecessor of i with out-degree $c^{\rm out}$ and in-degree $c^{\rm in}$ conditional on the out-degree. Similarly, the probability for a gene to belong to the GC in the cavity graph, where its *successor* μ has been removed, is found averaging (18)

$$\tilde{g} = \sum_{d^{\rm out}} P_d^{\rm out}(d^{\rm out}) \frac{d^{\rm out}}{\langle d^{\rm out}\rangle} \sum_{d^{\rm in}=1}^{\infty} P(d^{\rm in}|d^{\rm out}) \left[1 - \left(1-\tilde{t}\right)^{d^{\rm in}}\right]. \tag{B.2}$$

Under the assumption of independence of in- and out-degrees $P(d^{\rm in}|d^{\rm out}) = P_d^{\rm in}(d^{\rm in})$ and $P(c^{\rm in}|c^{\rm out}) = P_c^{\rm in}(c^{\rm in})$, one obtains

$$\tilde{g} = \sum_{d=1}^{\infty} P_d^{\rm in}(d) \left[1 - \left(1-\tilde{t}\right)^{d}\right] \tag{B.3}$$

$$\tilde{t} = \sum_{c=1}^{\infty} P_c^{\rm in}(c) \tilde{g}^{c}. \tag{B.4}$$

Therefore, the cavity probability $\tilde{t} = \langle n_\mu^{(i)}\rangle$ is equal, up to differences $\mathcal{O}(N^{-1})$, to $t = \langle n_\mu\rangle$ and the same holds true for $\tilde{g} = \langle n_i^{(\mu)}\rangle$ and $g = \langle n_i\rangle$. Upon introducing the generating functions $G^{(d)}(x) = \sum_k P_d^{\rm in}(k)x^k$, $G^{(c)}(x) = \sum_k P_c^{\rm in}(k)x^k$, the above equations take the compact form

$$\tilde{g} = 1 - G^{(d)}(1-\tilde{t}) \equiv f_1(\tilde{t},\tilde{g}) \tag{B.5}$$

$$\tilde{t} = G^{(c)}(\tilde{g}) \equiv f_2(\tilde{t},\tilde{g}). \tag{B.6}$$

Similarly, g and t are found from (20).

The point $(\tilde{g},\tilde{t}) = (0,0)$ is always a solution to these equations, corresponding to the situation where there is no GC in the network. Thus, the point at which this solution is no longer stable, will be the point at which a giant cluster emerges in the network. Denoting $\mathbf{x} = (\tilde{g},\tilde{t})^T$, and expanding the above equations around the fixed point, gives $\mathbf{x} = \mathbf{J}\mathbf{x}$, where $\mathbf{J}$ is the Jacobian evaluated at the fixed point

$$\mathbf{J} = \frac{\partial\left[f_1, f_2\right]}{\partial\left[\tilde{g},\tilde{t}\right]}\Big|_{(\tilde{g},\tilde{t})=(0,0)} \tag{B.7}$$

The solution $(\tilde{g},\tilde{t}) = (0,0)$ is stable provided that $|\mathbf{J}| < 1$. Taking partial derivatives of (B.5) and (B.6), one has

$$|\mathbf{J}| = \begin{vmatrix} 0 & \langle d^{\rm in}\rangle \\ P_c^{\rm in}(c=1) & 0 \end{vmatrix}. \tag{B.8}$$

Hence, for Poisson degree distributions, $P_d^{\rm in}(d) = \pi_{d_2}(d)$ and $P_c^{\rm in}(c) = \pi_{c_1}(c)$, a GC will exist in the network if

$$|-\alpha c_2 c_1 e^{-c_1}| \leq 1. \tag{B.9}$$

This gives the percolation threshold (21) for the non-linear dynamics of the bipartite Boolean network. For Poisson degree distributions, the averages in (B.5) and (B.6) can be calculated exactly by using $\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ and $d_2 = \alpha c_2$, to find

$$g = \tilde{g} = 1 - e^{-\alpha c_2 \tilde{t}}, \qquad t = \tilde{t} = e^{c_1(\tilde{g}-1)} - e^{-c_1}. \tag{B.10}$$

These curves are plotted in figure 4.

Appendix B.2. Linear dynamics

The percolation threshold for the linear dynamics can be found following a similar line of reasoning to that for the non-linear dynamics. The key difference is that there is no longer a hard constraint requiring that for a TF to belong to the GC, all the genes contributing to it must belong to it as well. Instead it is sufficient that at least one of the genes contributing to a TF belongs to the GC. Thus, the equations for the genes' and TFs' indicator variables become symmetric and one obtains for the cavity probabilities,

$$\tilde{g} = \sum_{d^{\rm out}} P_d^{\rm out}(d^{\rm out}) \frac{d^{\rm out}}{\langle d^{\rm out}\rangle} \sum_{d^{\rm in}=1}^{\infty} P(d^{\rm in}|d^{\rm out}) \left[1-\left(1-\tilde{t}\right)^{d^{\rm in}}\right] \tag{B.11}$$

$$\tilde{t} = \sum_{c^{\rm out}} P_c^{\rm in}(c^{\rm out}) \frac{c^{\rm out}}{\langle c^{\rm out}\rangle} \sum_{c^{\rm in}=1}^{\infty} P(c^{\rm in}|c^{\rm out}) \left[1-(1-\tilde{g})^{c^{\rm in}}\right] \tag{B.12}$$

which simplify, under the assumption of independence of in- and out-degrees, to

$$\tilde{g} = \sum_{d=1}^{\infty} P_d^{\rm in}(d) \left[1-\left(1-\tilde{t}\right)^{d}\right] \tag{B.13}$$

$$\tilde{t} = \sum_{c=1}^{\infty} P_c^{\rm in}(c) \left[1-(1-\tilde{g})^{c}\right]. \tag{B.14}$$

The point $(\tilde{g},\tilde{t})=(0,0)$ is still a solution. Evaluating the Jacobian at this point gives

$$\mathbf{J} = \begin{pmatrix} 0 & \langle d^{\rm in}\rangle \\ \langle c^{\rm in}\rangle & 0 \end{pmatrix} \tag{B.15}$$

which only depends on the *averages* $\langle d^{\rm in}\rangle = d_2$ and $\langle c^{\rm in}\rangle = c_1$ of the in-degree distributions. Hence, this gives $d_2 c_1 \geq 1$ as a general condition for the existence of a giant component, which is valid for *arbitrary* in-degree distributions $P_d^{\rm in}(d)$, $P_c^{\rm in}(c)$. Using conservation of links $d_2 = \alpha c_2$, this gives rise to the percolation threshold (22) for the linear dynamics. The probabilities that a gene or TF belong to the giant cluster, can again be calculated explicitely for the Poissonian case, giving

$$g = \tilde{g} = 1 - e^{-\alpha c_2 \tilde{t}}, \qquad t = \tilde{t} = 1 - e^{-c_1 \tilde{g}} \tag{B.16}$$

and have been plotted in figure 4.

Appendix C. Inferring the length of attractors

In this appendix, it is demonstrated how one can infer the length of a limit cycle attractor of the dynamics, from the overlap distributions in Fig. 6, obtained for simulations on a network with $(c1, c2, \epsilon) = (1, 10, 0.5)$. If we assume that the transient to the steady state is short on this network as generally demonstrated for increasing connectivity in figure 3, then the dynamics will converge to the limit cycle attractor in $\mathcal{O}(1)$ time steps. Next, if the time window Δt used to perform the average over the steady state is much greater than the length of the limit cycle ℓ, the limit cycle will be fully traversed $N_{\Delta t}$ times. There can also be a fraction of the limit cycle length x traversed in each Δt as well as the $N_{\Delta t}$ full cycles giving $\Delta t = \ell\,(N_{\Delta t} + x)$. Over many simulation runs the average of x would be expected to be $\langle x\rangle = 0.5$. If the probability that any gene in the network is expressed is given by a then the probability that a

Δt	$\approx \langle \Delta q \rangle$	ℓ
500	2.3×10^{-3}	4.6
$2,000$	6.3×10^{-4}	5.0
$4,500$	2.6×10^{-4}	4.7

Table C1. Length of limit cycle attractors inferred from mean deviation from $q_{\alpha\beta} = 1$ in distributions of the overlap using different time windows for the steady state averaging.

gene is expressed differently in a different simulation run β is given by $2a(1-a)$. Then the average deviation in the overlap from $q_{\alpha\beta} = 1$ is given by

$$\langle \Delta q \rangle = \frac{2a(1-a)x\ell}{\Delta t}\,, \tag{C.1}$$

where $x\ell$ is the number of sites on which there is a difference in gene expression between simulation runs. Using this expression for $\langle \Delta q \rangle$ and the mean values of the deviation from $q_{\alpha\beta} = 1$ in figure 6 one can infer the length of the limit cycle attractor for that network shown in the table C1. There is also an $\mathcal{O}(1/N)$ effect on the overlap of the initially introduced TFs hitting finite clusters that are able to sustain a net steady state gene expression level alongside the contribution from the giant cluster. Therefore, the network used to construct the overlap distributions in figure 6 likely has an attractor that is a limit cycle of length $\ell \simeq 5$. However, this periodicity is difficult to identify directly from trajectories of the dynamics.