Article

Fibroblast-specific protein-protein interactions for myocardial fibrosis from MetaCore network

Klaus M. Frahm [1], Ekaterina Kotelnikova[2], Oksana Kunduzova[3,4] and Dima L. Shepelyansky[1,*]

[1] Laboratoire de Physique Théorique, Université de Toulouse, CNRS, UPS, 31062 Toulouse, France
[2] Clarivate Analytics, Barcelona 08025, Spain
[3] National Institute of Health and Medical Research (INSERM) U1048, 31432, Toulouse, Cedex 4, France
[4] Université de Toulouse, UPS, 31062 Toulouse, France
* Correspondence: dima@irsamc.ups-tlse.fr; Tel. +33-56155-60-68 (D.L.S.)

Received: September 4, 2024; Accepted: date; Published: date

Abstract: Myocardial fibrosis is a major pathologic disorder associated with a multitude of cardiovascular diseases (CVD). The pathogenesis is complex and encompasses multiple molecular pathways. Integration of fibrosis-associated genes into the global MetaCore network of protein-protein interactions (PPI) offers opportunities to identify PPI with functional and therapeutic significance. Here, we report the generation of a fibrosis-focused PPI network and identification of fibroblast-specific PPI driving reparative and reactive myocardial fibrosis. In TGFb-mediated fibroblast activation, PPI hubs predict new regulatory mechanisms for fibrosis-associated genes. We introduce an efficient Erdös barrage approach to suppress activation of a number of fibrosis-associated nodes in order to reverse fibrotic cascades. Our results suggest that PPI prediction model can offer network insights into fibrosis mechanisms and can complement future experimental efforts to counteract cardiac fibrosis.

Keywords: fibrosis; Markov chains; Ising spin, Monte Carlo method, opinion formation, directed networks; protein-protein interactions

1. Introduction

Myocardial fibrosis is a major pathologic disorder associated with a multitude of cardiovascular diseases (CVD) [1]. In the heart, tissue fibrotic remodeling is characterized by abnormal fibroblast activation and excessive extracellular matrix (ECM) protein accumulation [2]. Although a number of factors have been implicated in orchestrating the fibrotic response, tissue fibrosis is dominated by a central mediator: transforming growth factor-β (TGF-β) [3]. Sustained TGF-β production leads to a continuous cycle of growth factor signaling and deregulated matrix turnover [4] . However, despite intensive research, the biomolecules that orchestrate fibrosis are still poorly understood and as a result, effective strategies for limiting fibrosis are lacking [2,4]. Considering the complex heterogeneity of fibrosis, research strategy on a system-level understanding of the disease using mathematical modeling approaches is a driving force to dissect the complex processes involved in fibrotic disorders. Recently, we have reproduced the classic hallmarks of aberrant cardiac fibroblast activation leading to fibrosis and provided a powerful toolbox for characterization of cardiac fibroblast activation [5]. Although the pathogenesis of fibrotic remodeling has not been well identified, accumulated evidence suggests that multiple genes/proteins and their interactions play important roles in disease scenario [6].

Fibrotic remodeling is a complex process caused by genetic abnormalities that alter protein-protein interactions (PPI) [5,7]. In the heart, PPI interfaces represent a highly promising, although challenging, class of potential targets for therapeutic options. In fibrosis, PPI form signaling nodes and hubs that transmit pathophysiological cues along molecular networks to achieve an integrated biological output, thereby promoting fibrogenesis and fibrosis progression. Thus, pathway perturbation, through disruption of PPI critical for fibrosis, offers a novel and effective strategy for curtailing the transmission

of pro-fibrotic signals. Deciphering of fibrosis-specific PPI would uncover new mechanisms of fibrotic signaling for therapeutic interrogation. Thus a mathematical analysis of the global PPI MetaCore network [8] can provide useful deep insights in the understanding of fibrosis progression processes.

In this work we use the PPI MetaCore network [8] to perform a mathematical modeling of fibrosis progression. The results reported in [5] determined the protein pro-fibrotic responses as a feedback on TGF protein stimulation. Thus these results [5] support the known fact that the TGF protein plays an important role in fibrosis tissue [3] and establish proteins with most positive and most negative response in cardiac fibroblasts.

A variety of biological applications of the MetaCore network are described in [9,10]. The general statistical properties of the MetaCore network are presented in [11]. The applications of Google matrix algorithms to the fibrosis PPI based on the MetaCore network are described in [12] using the fibrosis responses obtained in [5]. We note that the Google matrix algorithms [13–15] find a variety of useful applications in modern complex networks [16] including World Wide Web, Wikipedia, world trade etc.

In our opinion an important advantage of the MetaCore network is that it presents a global network structure of PPI with about 40000 proteins and important molecules. This allows to perform a mathematical modeling of fibrosis progression over the whole network. Thus in this work we present such a modeling which assumes that at an initial stage there is a certain number of proteins which activate a fibrosis progression of other proteins via their network links to other proteins. These initial proteins are always marked as red network nodes and assumed to be always red permanently generating fibrosis progression (thus being always red or with fixed Ising spin up with $\sigma_i = 1$). Among the other proteins, we assume that there is an initial group of proteins which inhibit fibrosis progression being of fixed blue color (thus being always blue or with Ising spin down with $\sigma_i = -1$). The remaining other proteins (or network nodes) are assumed to be in neutral state of white color or spin being zero ($\sigma_i = 0$). The fibrosis progression is modeled by the asynchronous Monte Carlo process when at each step a spin of a given node (which is not fixed) is determined as the total spin sign of other nodes linked to it with certain weights of Markov chain transitions. In a certain sense this rule corresponds to a situation when a person takes an average opinion of his network friends linked to him. After many of such Monte Carlo steps the system converges to a steady-state when all network spins obtain a fixed polarization being up (red or fibrosis activated) or down (blue or healthy protected) and some of the nodes may remain at zero neutral spin state (white) being practically disconnected from the red or blue network nodes.

We note that the above Monte Carlo steps describe a process of opinion formation and the global vote of nodes (society members) between red and blue options. For complex social networks [16] there are numerous studies of opinion formation, reviewed in [17]. Various interesting features of voter models and opinion formation on networks had been obtained and described in [17,19–24]. Recently it was argued that the opinion formation on the world trade network can be linked with country preference to trade in one or another currency (e.g. US dollar or hypothetical BRICS currency) [25]. The important new element appeared in [25] is that opinion of certain network nodes (countries) is considered to be fixed since it is assumed that they prefer to trade always with fixed currency of USD or BRICS. A further development of this approach with nodes of fixed opinion is done in [26] for Wikipedia networks when initially there are two groups of nodes of fixed red and blue opinions but all other nodes have neutral state (white opinion or spin zero) and their fixed opinion emerges in the result of asynchronous Monte Carlo process. This process is viewed as the Ising Network Opinion Formation (INOF) model. In fact this INOF model is directly suitable for the modeling of fibrosis progression in the PPI MetaCore network and we are using it here with certain slight modifications described in the next Section. We call this modified system as the Ising Network FIbrosis (INFI) model. It should be pointed out that a somewhat similar asynchronous Monte Carlo process is used in models of associative memory (see e.g. [27,28]).

In this work, using the results reported in [5,12], we take in the INFI model an initial state of 10 fixed red nodes that produce positive activation response of fibrosis on growth factor-β (TGF-β)

considering them as fixed 10 red protein nodes. We also consider 6 (or 10, 14) fixed blue protein nodes with negative response (see details in next Section). As a result of fibrosis progression via asynchronous Monte Carlo iterations we obtain the steady-state of affected and healthy protein nodes with about 60 - 80 percents of proteins affected by fibrosis. This corresponds to a strong fibrosis activation resolution of a large part of the system.

To reduce this strong fibrosis activation we develop an efficient barrage approach based on the analysis of Erdös proteins linked with fixed red nodes. We show that this Erdös barrage constructed in a clever manner allows to obtain a striking reduction of fibrosis activated resolution nodes by a factor of 300 so that almost all protein nodes are transformed in the healthy phase. We hope that the barrage strategy developed in this work will find applications in clinical experiments with fibrosis progression.

The paper constructing as follows: Section 2 describes the data sets, construction of Markov chain transitions, Google matrix and the INFI model, Section 3 presents the obtained results and Discussion and conclusion are given in Section 4. In the Appendix we provide additional data and Figures.

2. Data sets and INFI model description

2.1. Network data sets

In this work we use the same global PPI MetaCore network as in [12]. It contains $N = 40079$ nodes with $N_\ell = 292191$ links (without self connections which existed in [11]). The number of directed activation/inhibition links is $N_{\ell+}/N_{\ell-} = 65157/49321 \simeq 1.3$ and the number of neutral directed links is $N_{\ell n} = N - N_{\ell+} - N_{\ell-} = 177713$. Here we do not take into account the bi-functional activation/inhibition nature of links. Thus we simply have a directed network with N nodes and number of directed links being N_ℓ.

For convenience, we present in Table 1, taken from [12], 54 selected fibrosis proteins (nodes). These nodes are composed with 4 TGF-β proteins/nodes ($K_t = 1, 2, 3, 4$), 20 "up-proteins" ($K_u = 1, \ldots, 20$), 20 "down-proteins" ($K_d = 1, \ldots, 20$), both obtained from experiments [5] (as described above) and 10 new "X-proteins" (or "X-nodes"; $K_x = 1, \ldots, 10$) which have a significant influence on the other 44 nodes. The 4 TGF-β nodes correspond to different isoforms of this protein. As in [12] we have 4 groups of proteins and we use a specific index for each group: TGF-β proteins with index $K_t = 1, 2, 3, 4$; up-proteins with a strongest positive response noted by index $K_u = 1, \cdots, 20$ (ordered by the positive response with the strongest response for $K_u = 1$); down-proteins with a strongest negative response noted by index $K_d = 1, \cdots, 20$ (ordered by the modulus of negative response with the strongest response modulus for $K_d = 1$); external proteins noted by index K_x ordered by their local PageRank index (strongest PageRank probability of these 10 proteins is at $K_x = 1$; see more detail below). All these 54 proteins have their global index $K_g = 1, \cdots, 54$ used in Table 1.

Table 1. Table of the subset of $N_r = 54$ selected fibrosis proteins (nodes). Here K_g represents the global index of this group, $K_{t,u,d,x}$ represent the index of the four subgroups of 4 TGF-β proteins, 20 up-proteins, 20 down-proteins and 10 additional X-proteins; the K (K^*) index represents the PageRank (CheiRank) index for the global MetaCore network of $N = 40079$ nodes; the last column gives the associated protein names.

K_g	$K_{t,u,d,x}$	K	K^*	Protein
1	$K_t = 1$	10780	26299	TGF-β 0
2	$K_t = 2$	235	5690	TGF-β 1
3	$K_t = 3$	968	25073	TGF-β 2
4	$K_t = 4$	4726	29508	TGF-β 3
5	$K_u = 1$	28737	25928	ADAMTS16
6	$K_u = 2$	3478	25137	FGF21
7	$K_u = 3$	40048	28152	TNFSF18
8	$K_u = 4$	2467	19160	ACAN
9	$K_u = 5$	1489	24511	RPH3A
10	$K_u = 6$	26600	29559	ADAMTS8
11	$K_u = 7$	34769	39960	MEGF6
12	$K_u = 8$	26295	27326	SV2B
13	$K_u = 9$	27111	36021	C1QTNF3
14	$K_u = 10$	34616	39841	ANO4
15	$K_u = 11$	12696	16566	IL11
16	$K_u = 12$	26624	23640	CDH10
17	$K_u = 13$	7263	30243	HTR2B
18	$K_u = 14$	4647	6551	LAMA1-1
19	$K_u = 15$	8342	26295	LAMA1-2
20	$K_u = 16$	4021	8252	RAPGEF4
21	$K_u = 17$	29945	36964	DNER
22	$K_u = 18$	22159	8569	GALNT3
23	$K_u = 19$	29145	15531	ACSBG1
24	$K_u = 20$	24786	8735	OLFM2
25	$K_d = 1$	19039	28262	CLEC3B
26	$K_d = 2$	26477	28290	SCARA5
27	$K_d = 3$	26109	11185	SLC10A6
28	$K_d = 4$	6360	29204	CXCL5
29	$K_d = 5$	14952	8729	MYOC
30	$K_d = 6$	5961	22288	IFITM1
31	$K_d = 7$	5599	4483	ANGPTL4
32	$K_d = 8$	25538	17434	SELENBP1
33	$K_d = 9$	18938	33179	FMO1
34	$K_d = 10$	34080	39427	GPR88
35	$K_d = 11$	6276	22141	HMGCS2
36	$K_d = 12$	37060	28328	LGI2
37	$K_d = 13$	9162	2485	PTN
38	$K_d = 14$	513	5974	ADORA2A
39	$K_d = 15$	7789	22652	GFRA1
40	$K_d = 16$	6718	8844	IL1R2-1
41	$K_d = 17$	35446	28306	IL1R2-2
42	$K_d = 18$	12148	3444	PEG10
43	$K_d = 19$	27829	36195	FMO2
44	$K_d = 20$	1973	24994	COX4I2
45	$K_x = 1$	3	13	β-catenin
46	$K_x = 2$	4	6	p53
47	$K_x = 3$	11	10	ESR1
48	$K_x = 4$	13	25	STAT3
49	$K_x = 5$	22	11	RelA
50	$K_x = 6$	38	82	PPAR-γ
51	$K_x = 7$	111	767	IKK-β
52	$K_x = 8$	179	198	SNAIL1
53	$K_x = 9$	237	1520	MMP-14
54	$K_x = 10$	578	2123	Flotillin-1

2.2. Without formulas: methods, characteristics and expected network results

As in [12] here we present qualitative explanations without formulas of the mathematical methods and characteristics described in the next Subsections. Our aim here is to give a global view on our approach for a common reader.

We use the MetaCore directed network [8] which represents an action of a protein A on a protein B in a form of a directed link (edge) for $N = 40079$ proteins forming the network nodes (proteins). These links are obtained on the basis of careful and detailed analysis of scientific literature about thousands of experiments of various research groups that allowed to collect information about PPI and thus generated a network database with $N = 40079$ nodes and $N_\ell = 292191$ links. Certain medical applications of the MetaCore network can be found in [9,10].

The universal mathematical methods to analyze such networks are generic and based on the concept of Markov chains and Google matrix [13–15]. The validity of these methods has been confirmed for various directed networks from various fields of science. Therefore, since the Google matrix analysis is based on a generic mathematical foundation, we expect that this analysis will also work efficiently for PPI networks.

The Google matrix of the global MetaCore PPI network G is constructed with specific rules briefly described in the next Subsection 2.3 and in detail in [13–15]. The matrix G is obtained from a matrix of Markov chain transition elements S_{ij} that give weights of transitions between nodes. The important property of G is that its application (multiplication) to an arbitrary initial vector v preserves the probability and the normalization of this vector (sum of all vector elements) remains constant (taken to be unity). As a result of multiple multiplications by G any initial vector converges to the steady-state distribution given by the PageRank vector P. The components of this vector represent the probabilities of each node (protein) in this limit. The nodes with the highest probabilities are the most influential nodes of the network (all nodes are monotonically ordered by decreasing values of the PageRank components which provides the "PageRank index" K such $K(j) = 1, 2, \ldots$ for nodes j with largest values $P(j)$). These nodes have typically many ingoing links and it is likely that some of these ingoing links come from other nodes that also have large PageRank values.

For the inverse network, in which all link directions are inverted, the corresponding PageRank is called CheiRank vector P^* [15]. The highest probabilities $P^*(j)$ are for nodes j with the CheiRank index $K^*(j) = 1, 2, \ldots$ being the most communicative nodes. They typically have many outgoing links.

In the INFI model Ising spins with polarization ± 1 (red or blue) or 0 (white) are placed at each node of network. They describe fibrosis activated protein state (red or $\sigma_j = +1$) which can activate other proteins, healthy protein state (blue or $\sigma_j = -1$) which can prepare other proteins and neutral state proteins (white or $\sigma_j = 0$) that cannot affect other proteins but can be activated or repaired by red or blue proteins. The initial spin configuration has certain fixed red proteins with index $K_g = 1, \ldots, 10$ (see Table 1) and fixed blue proteins with index $K_g = 25, \ldots, 30$ (or 34 or 38) whose colors always remain unchanged, all other proteins are white (or $\sigma_j = 0$). These white proteins, and also red or blue non-fixed proteins, can change their color depending on the majority color of nodes directly linked to it (taking into account the weights of ingoing and outgoing links of the matrix S_{ij} taken without dangling nodes). This choice is similar to social relations when a member of society takes his opinion to be imposed by opinions of other members related with him. An asynchronous Monte Carlo process performs multiple changes of spins leading to the convergence to a steady-state polarization of all spins. The fraction of red spins f_r determines if the fibrosis progression affected a major part of network or not. We give the mathematical definitions of the INFI model in next Subsections.

2.3. Markov chains, Google matrix, PageRank and CheiRank

In this section we remind some basic mathematical definitions of Markov chains and how the Google matrix and the related PageRank and CheiRank vectors for the MetaCore network are constructed. For this we use a simple version of the MetaCore network where initial links are not weighted (see Refs. [11,12] for some details about a more refined version with bi-functional links).

First, one introduces the adjacency matrix with elements $A_{ij} = 1$ if a node j points to another node i (different from j) and $A_{ij} = 0$ if there is no link from j to i or if $i = j$ (we use the same convention as in Ref. [12] where all diagonal elements are chosen to be $A_{jj} = 0$ even if there is a link from a node j to itself). Then we define a matrix $\tilde{S}$ by $\tilde{S}_{ij} = A_{ij}/\sum_l A_{lj}$ if $\sum_l A_{lj} \neq 0$ and $\tilde{S}_{ij} = 0$ otherwise (i.e. for columns j where all elements A_{lj} are 0). This matrix does not yet describe a Markov process since it has potentially some zero columns j (such nodes j are also called dangling nodes). However, in this work we will use a Monte Carlo process based on this particular matrix (see next section).

To obtain a proper stochastic matrix S one replaces $S_{ij} = 1/N$ for the dangling nodes j and $S_{ij} = \tilde{S}_{ij}$ for the other nodes (dangling nodes have no outgoing links so that a column of dangling node j has only zero elements in the matrix $\tilde{S}_{ij}$). Here S_{ij} represents the transition probability for a random surfer from node j to i and the column sum normalization $\sum_i S_{ij} = 1$ ensures the conservation of probability.

The Google matrix elements G_{ij} are typically defined by

$$G_{ij} = \alpha S_{ij} + (1 - \alpha)/N \tag{1}$$

where $\alpha = 0.85$ is the usual damping factor [13,14]. The Google matrix is also a proper stochastic matrix (column sum normalized) and here the random surfer jumps with probability α on the network according to S and with a probability $(1 - \alpha)$ to an arbitrary random node of the network. The damping factor modification helps to avoid possible isolated communities and ensures that the Markov process converges for long times rather quickly to a uniform stationary probability distribution. The latter defines the PageRank vector P which is actually the right eigenvector of the Google matrix G corresponding to the leading eigenvalue $\lambda = 1$, i.e. $GP = P$. The elements $P(j)$ of the PageRank vector correspond to the probability to find the random surfer on the node j in the stationary limit of the Markov process. The PageRank index $K(j)$ is obtained by ordering the nodes with decreasing values of $P(j)$, i.e. the highest (lowest) PageRank probability $P(j)$ corresponds to $K(j) = 1$ ($K(j) = N$). In this work we will use the K-rank of a node to characterize this node in a unique way. The PageRank probability $P(j)$ is typically related to the number of ingoing links pointing to node j. However, it also takes into account the "importance" (i.e. PageRank probability) of the nodes having a direct link to j.

One can also consider the network obtained by the inversion of all link directions. The same construction as described above provides a Google matrix noted as G^* and the corresponding PageRank vector is called the CheiRank vector P^*, defined by $G^*P^* = P^*$. The CheiRank probability $P^*(j)$ is typically related to the number of outgoing links weighted by the value $P^*(i)$ of nodes i having a link from j to i. The CheiRank index $K^*(j)$ is obtained by ordering the CheiRank vector with decreasing values of $P^*(j)$.

For further details about the properties of Google matrix, PageRank and CheiRank vectors, with many different example networks (also with proteins networks), we refer to [11,12,14,15].

2.4. Ising spin network, Monte Carlo process for INFI model

In this work, we consider an Ising type of model using the symmetrized matrix $\tilde{S} + \tilde{S}^T$ to model the effective spin interactions. We denote by σ_i the spin associated to the node (protein) i but here we allow for three different values being $\sigma_i = +1$ ("red state", "fibrosis state"), $\sigma_i = -1$ ("blue state", "healthy state") and also $\sigma = 0$ ("white state", "undetermined or neutral state"). We study this model using a specific Monte Carlo process similar to [26]. The white state will only be used in the initial condition and is supposed to be only temporary.

First, we fix 10 specific nodes permanently to the red state, those with $K_g = 1, \ldots, K_g = 10$ in the set of Table 1 corresponding to the four TGF proteins and the first 6 K_u-proteins with strongest positive response on the TGF stimulation [12]. We also fix n_b nodes permanently to the blue state being the nodes with $K_g = 25, \ldots, K_g = 24 + n_b$ corresponding to the first n_b K_d-proteins with strongest negative TGF-response. Here we choose mostly $n_b = 6$ (same number as the 6 K_u proteins with fixed red state)

but we also present some data for the cases $n_g = 10$ (same total number of fixed red proteins) and $n_g = 14$. Therefore $10 + n_b$ nodes have already a permanently fixed spin value $\sigma_i = \pm 1$.

We implement a Monte Carlo procedure that modifies the other $N_v = N - 10 - n_b$ remaining nodes with "variable" spin values (non fixed). For this at given iteration time τ, we update the N_v variable nodes by computing the sum over all nodes j of the network (fixed and variable) :

$$Z_i = \sum_{j(\neq i)} (\tilde{S}_{ij} + \tilde{S}_{ji})\sigma_j \tag{2}$$

and choosing $\sigma_i(\tau + 1) = +1$ if $Z_i > 0$, $\sigma_i(\tau + 1) = -1$ if $Z_i < 0$ and $\sigma_i(\tau + 1) = \sigma_i(\tau)$ (unchanged) if $Z_i = 0$. Without going into technical details, we mention that the numerical implementation of (2) exploits of course the sparse structure of the matrix $\tilde{S}_{ij}$. This update is done for every node i in the set of N_v variable nodes in a random order using a random permutation (thus no repetitions) which is chosen at the beginning of the iteration process. Once a spin value for a node i has been updated, the new value $\sigma_i(\tau + 1)$ will be used for the update of the subsequent nodes $\tilde{i}$. We note that somewhat similar asynchronous Monte Carlo process what used for bio-networks in [29,30] but there the matrix elements $\tilde{S}_{ij} + \tilde{S}_{ji}$ had values ± 1 and network sizes were relatively small (about 100).

This full update procedure for one time step and the full set of N_v variable nodes is repeated up to the iteration time $\tau_{\max} = 100$ (with different random permutations for the update order at each time step). We also verify if at any given value of τ there is ideal convergence where the spins no longer change. (This happens typically at $\tau_{last} \approx 10$; see below for details).

The full iteration procedure with $\tau_{\max} = 100$ time steps is repeated $R = 100000$ times with different random permutations and eventually also different random positions of initial blue or white nodes (depending on the precise type of initial condition).

During this procedure, we compute the fraction of red outcome $f_r(i)$ for each node as the number of times the node i has red state at $\tau_{\max}$ (at different realizations) divided over R and also the overall average $f_r = \sum_i f_r(i)/N_v$ over the nearly full network of variable nodes. To test the convergence, we also compute the overall average at intermediate values of τ.

In order to test the procedure and also in the direct influence of the permanently fixed nodes, we first choose an initial condition where all N_v variable nodes are initialized to the white state. Figure A1 illustrates for this case the convergence of (the overall network average) $f_r(\tau)$ with iteration time τ. The convergence seems to be very good at $\tau > 5$ with nearly constant values on graphical precision. However, a closer look at the data shows that $|f_r(\tau + 1) - f_r(\tau)| \sim 10^{-5}$ for $\tau = 10$ and the value of $f_r(\tau)$ becomes constant only at $\tau \geq \tau_{last} \approx 70$ with typical last non-vanishing differences $|f_r(\tau_{last} - 1) - f(\tau_{last})| \sim 10^{-10}$ and $|f_r(\tau_{last} + 1) - f(\tau_{last})| = 0$. We have verified that this behavior is due to the fact that typically at $\tau \approx 10$ there is exact convergence for a given realization of the random permutations (random pathways) with no further modifications of the spins at $\tau > 10$. However, there are rare realizations with larger values, up to $\tau = \tau_{last} \approx 70$ for exact convergence.

In the next section, we present the results for different quantities and also other initial conditions.

3. Results

3.1. Numerical results

All results presented in this work were obtained with maximal iteration time $\tau_{max} = 100$ and $R = 100000$ realizations for (potential) different initial conditions and random permutations for the Monte Carlo procedure. Also when we describe a particular initial condition it is implicitly understood that the permanently fixed 10 red and n_b blue nodes given above are indeed fixed to these values from the very beginning, i.e. if we say "that there are n_{ib} initial blue values" (with $n_{ib} = 0, 1, \ldots$) we mean that there are n_{ib} blue nodes *on the set of N_v variable (non-fixed) nodes*, other variable nodes are initialized to white values and the fixed nodes have still their 10 red and n_b blue values (in particular the value of n_{ib} does not include the number of n_b fixed blue nodes).

In Figure 1, we show for the case of white or zero initial spins (for the N_v variable nodes; i.e. $n_{ib} = 0$) and the three cases of $n_b = 6, 10, 14$ initial blue fixed nodes (and 10 fixed nodes as explained above) the probability distribution (normalized by an integral) of fraction of red outcomes $f_r(i)$ for the nodes obtained by a histogram with bin width $\Delta f_r = 0.01$. These distributions are strongly peaked at values close to $f_{r,peak} \approx 0.93$ ($n_b = 6$), $f_{r,peak} \approx 0.81$ ($n_b = 10$), $f_{r,peak} \approx 0.65$ ($n_b = 14$) for roughly 82 % of nodes. The secondary peak at $f_r = 0$ corresponds to the fraction (number) ≈ 0.18 (7230) of nodes (same value for the three cases of n_b) which stay white after 100 iterations for all $R = 100000$ pathway realizations. This set of stable white nodes correspond to nodes not connected to the small number of fixed red and blue nodes. However, despite the small value of only 10 initial fixed red notes the majority of the other nodes has a red outcome, especially for $n_b = 6$ and with slightly reduced values for $n_b = 10, 14$. Note that the bin width $\Delta f_r = 0.01$ in Figure 1 is still the rather large and histogram computations with $\Delta f_r = 0.001$ and $\Delta f_r = 0.0001$ show that the actual peaks of the distributions are much sharper as visible in Figure 1.

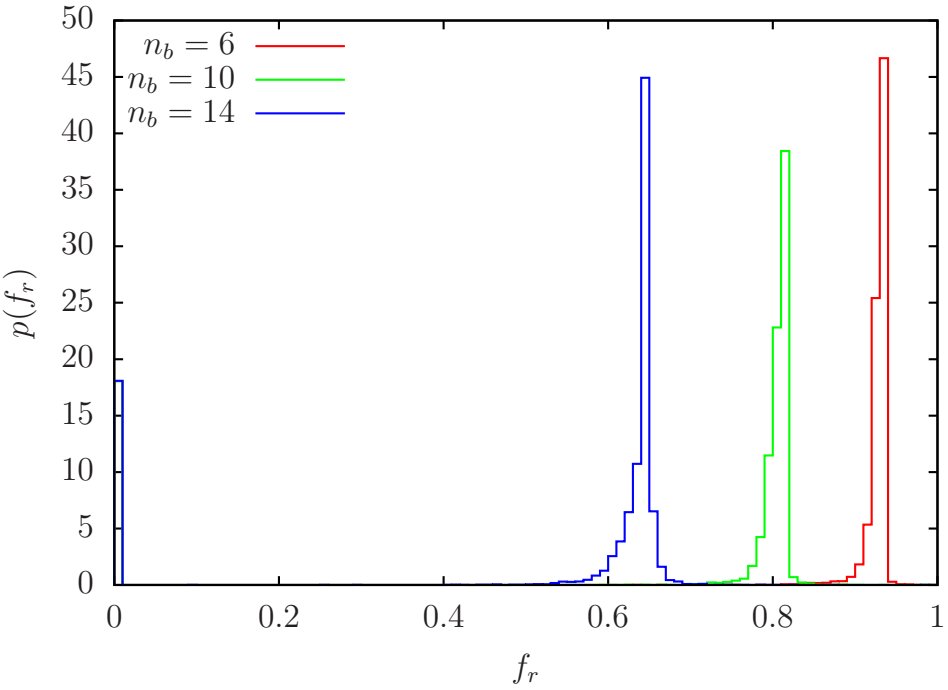


Figure 1. Probability density $p(f_r)$ of f_r for $n_b = 6, 10, 14$ obtained from a histogram with bin width $\Delta f_r = 0.01$ using $N_v = N - 10 - n_b$ data points $f_r(i)$ corresponding to the number of variable nodes. For each node i the value $f_r(i)$ is obtained as the fraction of red outcome ($\sigma_i = +1$) of $R = 100000$ pathway realizations for this node using the initial condition with no initial (variable) blue nodes, i.e. $n_{ib} = 0$. The distributions are normalized by $\int_0^1 p(f_r)\, df_r$ and there are strongly peaked at values close to $f_{r,peak} \approx 0.93$ ($n_b = 6$), $f_{r,peak} \approx 0.81$ ($n_b = 10$), $f_{r,peak} \approx 0.65$ ($n_b = 14$) for roughly 82 % of nodes. The secondary peak at $f_r = 0$ corresponds to the fraction ≈ 0.180 of nodes (same value for the three values of n_b) which stay white after 100 iterations for all $R = 100000$ pathway realizations.

It seems that with the white initial condition (except for the fixed nodes) nearly all nodes (except those in the stable white set) have a large probability for red outcome. Therefore, we try to reduce this red outcome by choosing a certain number n_{ib} of initial blue (variable) nodes and other nodes with white initial values. The question is also where to place these blue nodes. In a first model we choose the n_{ib} blue nodes at different random positions (for each of the $R = 100000$ pathway realizations) on the full set of variable nodes giving a set of $N_r = N_v$ of possible blue initial nodes with possible values $n_{ib} = 0, \ldots, N_r$ and a corresponding fraction $f_{ib} = n_{ib}/N_r$. Figure 2 shows for the case $n_b = 6$ (the results for $n_b = 10$ and $n_b = 14$ are very similar) the overall network average f_r at maximum iteration time as a function of f_{ib} (or n_{ib}). The initial value $f_r \approx 0.75855$ at $n_{ib} = 0$ corresponds to

the average which can also be obtained from the histogram data shown in Figure 1 which is roughly $0 \times 0.18 + 0.93 \times 0.82 \approx 0.76$. With increasing f_{ib} (or n_{ib}) the value of f_r decreases, e.g. $f_r \approx 0.1$ for $n_{ib} \approx 60$ and $f_r \approx 10^{-2}$ for $n_{ib} \approx 100$. However, this decrease is not optimal since the initial n_{ib} blue nodes can be on arbitrary random positions of available variable nodes. Therefore, we also try different network subsets for the potential initial blue nodes.

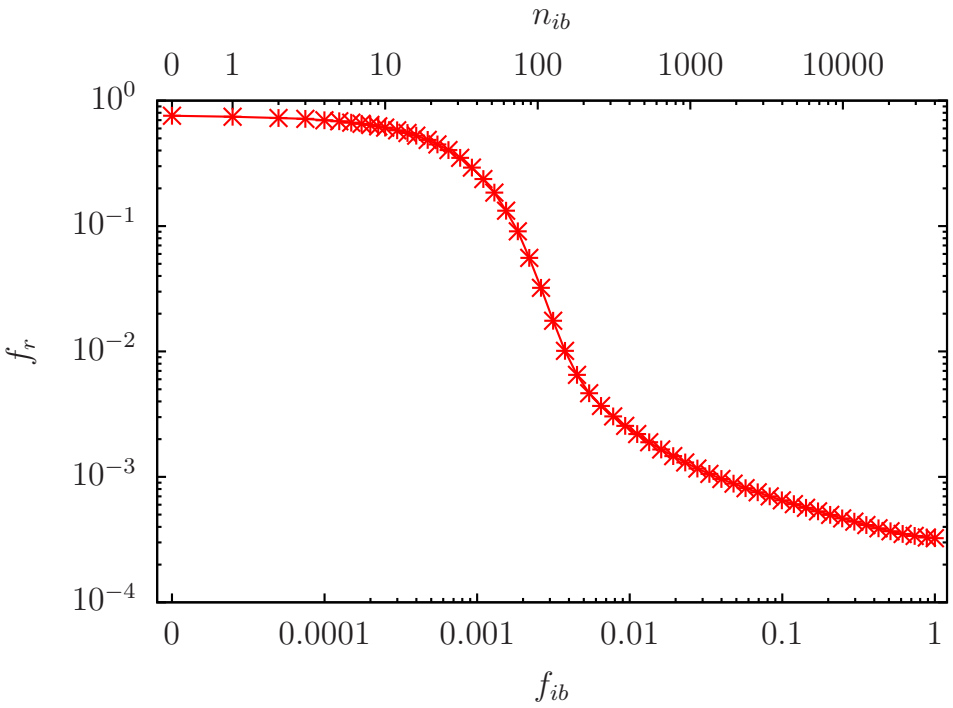


Figure 2. Probability of red outcome $f_r = \sum_i f_r(i)/N_v$ averaged over all (variable) N_v network nodes versus fraction f_{ib} of random initial blue nodes in the full set of $N_r = N_v$ variables nodes and for the case $n_b = 6$. The top x-axis shows the number $n_{ib} = N_r f_{ib}$ of initial (variable) blue nodes. For each value of n_{ib} the initial condition corresponds to n_{ib} initial blue nodes ($\sigma_i(\tau = 0) = -1$) with different random positions in the set of size N_r (for the $R = 100000$ pathway realizations). The representation is logarithmic on both axis except for the first data point at $n_{ib} = 0$ which has artificially been placed at a finite position below $n_{ib} = 1$ for practical reasons. Note that the value $f_r(n_{ib} = 0) \approx 0.75855$ of this data point can also be obtained from the distribution average $\int_0^1 f_r\, p(f_r)\, df_r$ from the data of Figure 1 for the case $n_b = 6$.

In Figure 3, we show the dependence of f_r on f_{ib} (or n_{ib}) for a reduced subset of $N_r = 38$ nodes corresponding to the 38 proteins of Table 1 which are not used for fixed red/blue values, i.e. $K_g = 11, \ldots, K_g = 24$ and $K_g = 31, \ldots, K_g = 54$. Now, n_{ib} represents the number of initial blue nodes on random positions in this subset (and white initial nodes on every other variable node). Now, the decrease of f_r with n_{ib} seems somewhat stronger, i.e. $f_r \approx 0.1$ for $n_{ib} \approx 8$ and $f_r \approx 10^{-2}$ for $n_{ib} \approx 16$, and we may conclude that this subset is more effective to reduce the red outcome ("to block fibrosis") than the full set of available variable nodes.

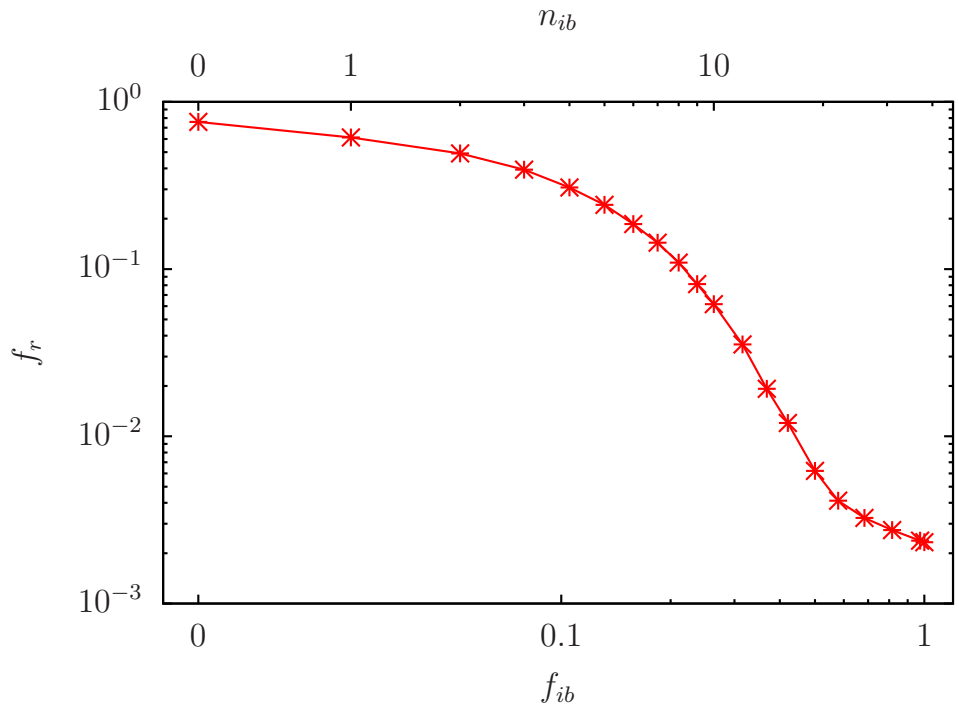


Figure 3. As Figure 2 for $n_b = 6$ but using a reduced set of potentially initial blue nodes with $N_r = 38$ using the non-fixed nodes of the set given in Table 1 (i.e. the nodes with $K_g = 11, \ldots 24$ and $K_g = 31, \ldots, 54$). For each value of n_{ib} the initial condition corresponds to n_{ib} initial blue nodes ($\sigma_i(\tau = 0) = -1$) in this reduced set of 38 nodes (with random positions for each of the $R = 100000$ pathway realizations). All other variable nodes have initial white values ($\sigma_i(\tau = 0) = 0$). The first data point at $n_{ib} = 0$ has the same value as in Figure 2 and it has also artificially been placed at a finite position below $n_{ib} = 1$ for practical reasons.

To determine a still more effective subset, we determine all nodes which have a direct link or inverse link to one of the 10 fixed red nodes (those with $K_g = 1, \ldots, K_g = 10$ in Table 1). This defines a particular set, of size $N_E = 353$, with Erdös number being unity with respect to the 10 fixed red nodes as HUB and using the symmetrized link matrix $\tilde{S}_{ij} + \tilde{S}_{ji}$. Figure 4 shows the dependence of f_r on f_{ib} (or n_{ib}) for this Erdös set with $N_r = N_E$. Now the decrease is even more effective with $f_r \approx 0.1$ for $n_{ib} \approx 6$ and $f_r \approx 10^{-2}$ for $n_{ib} \approx 14$,

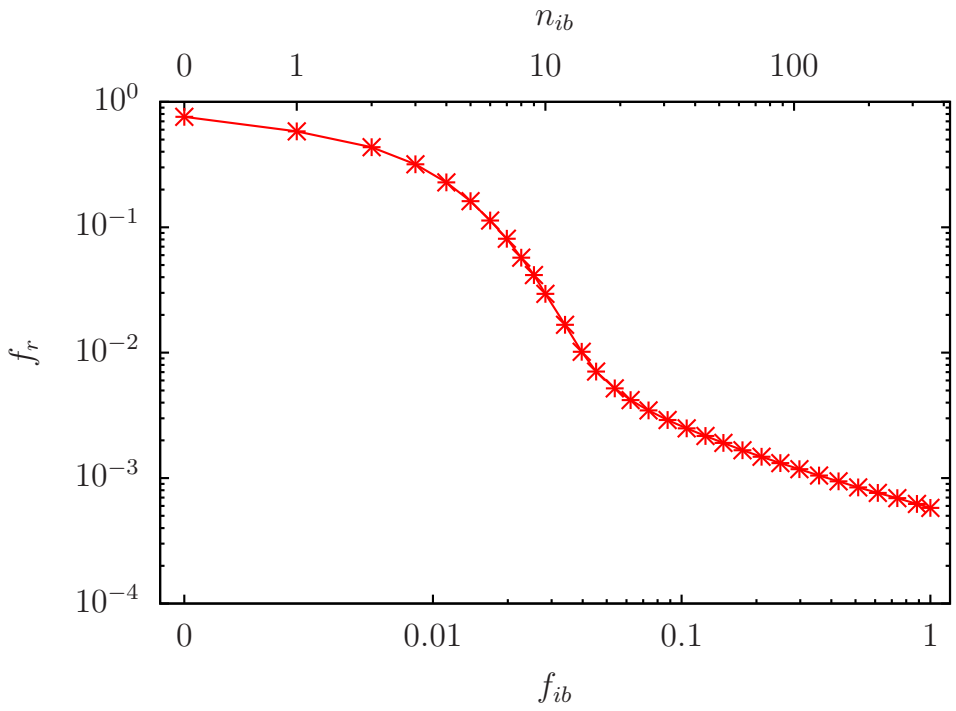


Figure 4. As Figure 3 for $n_b = 6$ but using the Erdös set (all nodes with direct links in both directions to the 10 fixed red nodes) as reduced set of potentially initial blue nodes with $N_r = N_E = 353$. The first data point at $n_{ib} = 0$ has the same value as in Figure 2 and it has also artificially been placed at a finite position below $n_{ib} = 1$ for practical reasons.

In Figure 5, we show in a color plot the dependence of $f_r(i)$ for the 54 nodes i belonging to the set of Table 1 on the index n_g which is a monotonic function of n_{ib} (essentially linear for $n_{ib} \leq 10$ and logarithmic for $n_{ib} > 10$; see caption of Figure 5 and Appendix Figure A1 for details). The data of Figure 5 correspond to the data of Figure 4 using the Erdös set with $N_E = 353$ as potential initial blue nodes (i.e. for each value of $n_{ib} \in \{0, \ldots, 353\}$ we have n_{ib} initial blue nodes at random positions in this set). In Fig 5, we can of course identify the 10 fixed red nodes ($K_g = 1, \ldots, 10$) and the $n_b = 6$ fixed blue nodes ($K_g = 25, \ldots, 30$) with either $f_r(i) = 1$ or $f_r(i) = 0$ respectively for all values of n_g. The other nodes follow quite closely the decrease of Figure 4 for the global average of f_r. However, for certain specific nodes ($K_g = 17, 23, 36, 38, 41, 44$) the decay of f_r with n_g (or n_{ib}) is less pronounced for $n_g > 10$. Apparently, these nodes are less likely to be blocked by a modest number of initial blue nodes.

We have also analyzed the data of Figure 2 (using all N_v variable nodes as potential initial blue nodes) and Figure 3 (using the remaining set of 38 non-fixed nodes of Table 1 as potential initial blue nodes) with similar color plots and in both cases we observe the same qualitative behavior as in Figure 5 : identification of fixed red/blue nodes, similar decrease of $f_r(i)$ with increasing n_{ib} for the other nodes and less pronounced decrease for the 6 specific nodes mentioned above.

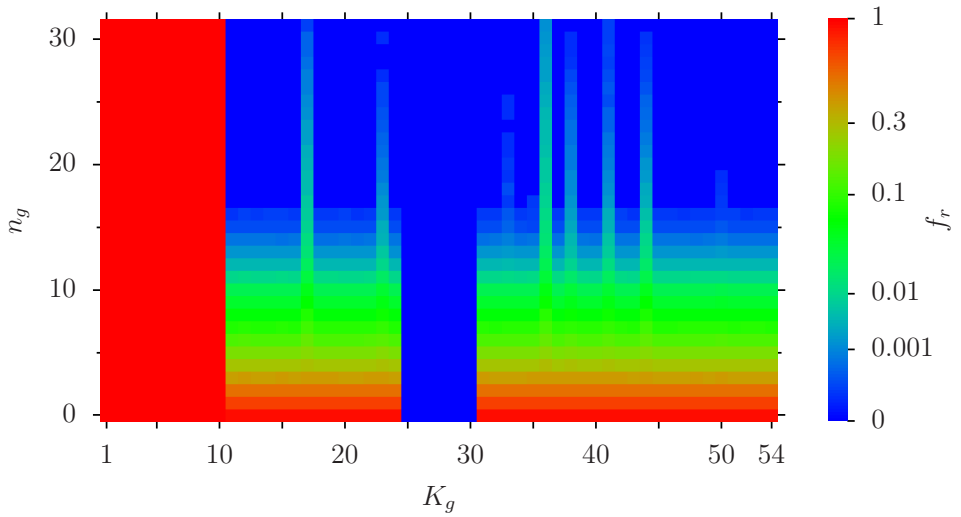


Figure 5. Color plot of $f_r(i)$ for the 54 nodes of the set of Table 1 for the case of Figure 4 (i.e. $n_b = 6$, using the Erdös group with $N_r = 353$ nodes as potential initial blue nodes and n_{ib} being the number of initial blue nodes at random positions in the Erdös group). The x-axis corresponds to the index K_g of Table 1 and the y-axis represents to the coarse-grained index $n_g = 0, \ldots, 31$ which corresponds to $n_{ib} = n_g$ for $n_g \leq 10$ (linear scale) and $n_{ib} \approx 10 \times 1.2^{n_g - 10}$ for $n_g > 10$ (logarithmic scale). See also Figure A2 which shows the link between n_{ib} and n_g. The values of the color bar correspond to $f_r(i)$ (i.e. red for $f_r(i) = 1$, green for $f_r(i) \approx 0.06$ and blue for $f_r(i) = 0$). Here small $f_r(i)$ values have been amplified to improve the visibility (non-linear scale in the Colombo).

The question arises which of the nodes of the Erdös set, or more generally, which configurations of few selected nodes of this set, are most effective to reduce the red outcome when selected as initial blue nodes. To answer this question, we compute for each node i of the Erdös set the red outcome $f_{rc}(i)$ averaged over the full network when this specific node i is selected as single initial blue node, i.e. with $n_{ib} = 1$ but now with different cases of given fixed positions (instead of random positions). Note that this quantity is different from $f_r(i)$ used in the histogram of Figure 1 which is the probability of red outcome of node i (not averaged over the network) with full white initial condition. The nodes of the Erdös set can be ordered with increasing values of the new quantity $f_{rc}(i)$ which provides a specific ranking index K_{f_r} in this set. In Table 2, we present in K_{f_r}-order the nodes (proteins) of the Erdös set which have either $K \leq 40$ or $K^* \leq 40$, either low K- or K^*-rank. It turns out these nodes also provide the lowest values of f_{rc} and K_{f_r} (with some holes for $K_{f_r} > 17$).

Table 2. Table of selected fibrosis proteins (nodes) belonging to a smaller subset of the Erdös group such that either $K \leq 40$ or $K^* \leq 40$ where the K (K^*) index represents the PageRank (CheiRank) index for the global MetaCore network of $N = 40079$ nodes. f_{rc} represents the average fraction of red nodes obtained when using the corresponding node as single initial non-fixed blue node and the index K_{f_r} is obtained by ordering the Erdös group of 353 elements with increasing values of f_{rc}; the last column gives the associated protein names. The shown subset of the Erdös group corresponds essentially to the nodes with lowest f_{rc}-value (up to $K_{f_r} = 17$). The four nodes marked by "$*$" in the first column are the nodes used for the subsequent example computations using 4 specific optimal nodes.

	K_{f_r}	K	K^*	f_{rc}	Protein
*	1	17	1	0.010691	c-Myc
*	2	4	6	0.029542	p53
*	3	342	7	0.038899	c-Fos
	4	14	14	0.053526	Androgen receptor
	5	22	11	0.075088	RelA (p65 NF-kB subunit)
	6	94	34	0.07674	HDAC1
	7	188	19	0.078157	p300
	8	6272	3	0.078989	IGF2BP3
	9	232	9	0.081749	SP1
	10	64	29	0.083725	HIF1A
*	11	3	13	0.083939	β-catenin
	12	203	21	0.13861	E2F1
	13	728	5	0.14694	SOX9
	14	432	24	0.14808	BRG1
	15	8	115	0.15027	EGFR
	16	72	18	0.15149	EZH2
	17	13	25	0.15649	STAT3
	20	480	40	0.17501	C/EBPβ
	21	38	82	0.20211	PPAR-γ
	24	298	33	0.20997	ELAVL1 (HuR)
	26	394	37	0.22213	CREB1
	31	6370	17	0.28502	PUM2
	32	17711	8	0.29183	CUX1 (p110)

Then we choose as example four "optimal" nodes with $K = 3, 4, 17, 342$ (marked with an asterix in the first column of Table 2). The nodes $K = 4, 17, 342$ occupy indeed the top three places in the K_{f_r}-rank while the node $K = 3$ "only" corresponds to the position $K_{f_r} = 11$. The reason for this choice is related to the fact that these four nodes are more uniformly optimal if we also choose a small number of $n_{ib} = 2, \ldots, 10$ initial blue nodes. In this case, using the data of Figure 4 (or the code to produce these data) it is actually possible to compute the conditional probability $f_{rc}(i)$ of a red outcome when the node i is by chance selected by the random initial condition for $n_{ib} > 1$ (as one of the initial n_{ib} blue nodes). For example for $n_{ib} = 4$ with $R = 100000$ different random initial conditions of 4 blues nodes out of 353 we have typically a bit more than 1000 realizations where an arbitrary fixed node i belongs to the random set of 4 initial nodes. This provides enough data for a reasonable average to compute the conditional probability of red outcome of the node i. In this way, it is possible to compute more general K_{f_r}-rankings as in Table 2, also for modest values of $n_{ib} > 1$. It turns out that these rankings produce roughly the same sets of nodes in the first places (with possible permutations between different n_{ib}-values) and the four selected nodes $K = 3, 4, 17, 342$ are indeed optimal as a group for the three values $n_{ib} = 2, 3, 4$ (and also some larger values).

In Figure 6, we present (for $n_b = 6$) the dependence of f_r on n_i for this small optimal set of 4 possible initial blue nodes (red square data points). Here the decrease of f_r with increasing n_i is indeed very strong with $f_r \approx 0.035$ for $n_{ib} = 1$ and $f_r \approx 0.00318$ for $n_{ib} = 4$. Furthermore, Figure 6 also provides the conditional probabilities of red outcome $f_{rc}([K_1, \ldots, K_{n_{ib}}]$ (black small circle data points) for specific configurations $[K_1, \ldots, K_{n_{ib}}]$ with one configuration for $n_{ib} = 0, 4$, four configurations for $n_{ib} = 1, 3$ and six configurations for $n_{ib} = 2$. We see that there are considerable fluctuations between

the different configurations in the red outcome for $n_{ib} = 1$ and $n_{ib} = 2$ with optimal values being $f_{rc}([17]) \approx 0.009477$ and $f_{rc}([3,17]) \approx 0.004863$. Appendix Figure A3 provides two panels of similar figures for the other cases $n_b = 10$ and $n_b = 14$.

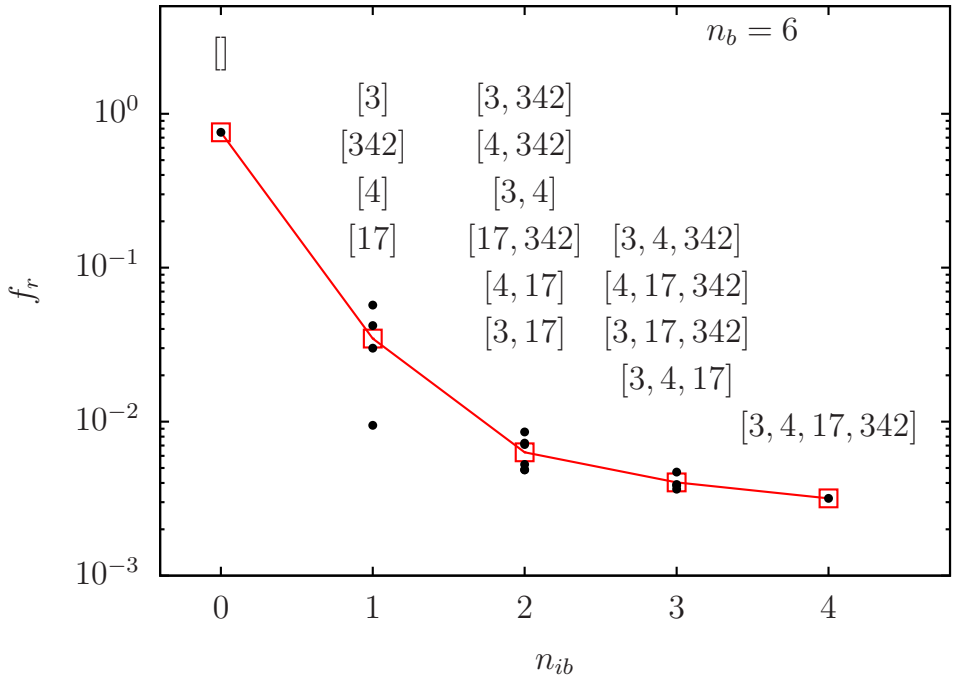


Figure 6. As Figs. 3 and 4 but with linear x-axis and using the optimal set with four K-values $[3,4,17,342]$ for potential initial blue nodes (i.e. $n_b = 6$, $N_r = 4$ and $n_{ib} = 0,\ldots,4$ being the number of initial blue nodes at certain positions in this set). The red square data points represent the average f_r with respect to all possible configurations of n_{ib} initial blue nodes and the black data points represent the conditional average $f_{rc}([K_1,\ldots,K_{n_{ib}}])$ for particular configurations $[K_1,\ldots,K_{n_{ib}}]$ of initial (variable) blue nodes (with K_j being K-rank values of nodes and $n_{ib} = 0,\ldots,4$). For each column the top (bottom) black data point corresponds to the top (bottom) configuration shown above. For $n_{ib} = 0$ ("empty" configuration "[]" with no initial variable blue node) and $n_{ib} = 4$ (full configuration "$[3,4,17,342]$" of all four nodes) there is only one configuration and therefore only one associated black data point.

Figure 7 is a similar color plot as Figure 5 but for the specific set of 4 optimal nodes $[3,4,17,342]$. The conclusions are similar to Figure 5 (confirmation of fixed red/blue nodes, strong decrease with increasing n_{ib} for other nodes and certain specific nodes with less pronounced decrease.

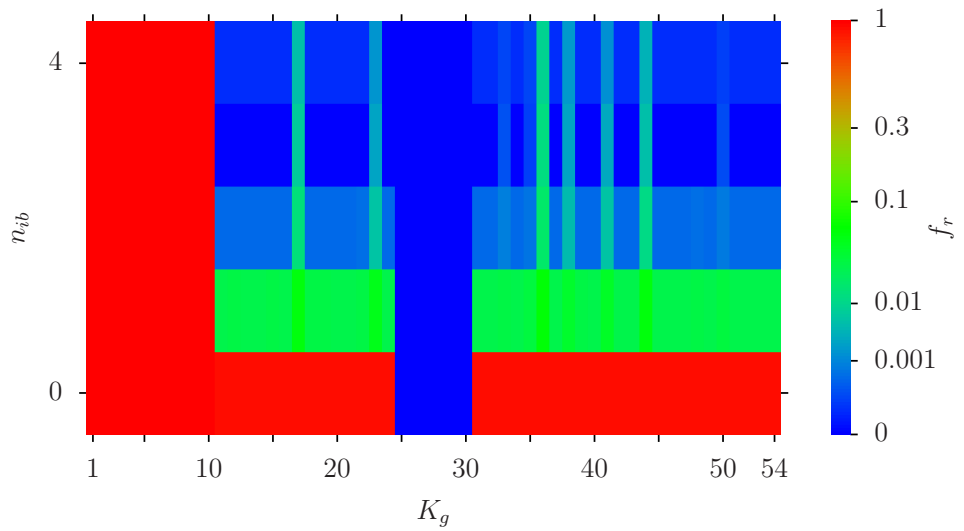


Figure 7. Color plot of $f_r(i)$ as in Figure 5 but using the optimal set with four K-values $[3, 4, 17, 342]$ for potential initial blue nodes (i.e. $n_b = 6$, $N_r = 4$ and $n_{ib} = 0, \ldots, 4$ being the number of initial blue nodes at random positions in this set). Here the y-axis corresponds directly to n_{ib} and the x-axis to K_g of Table 1. The values of the color bar correspond to $f_r(i)$ (with amplified scale as in Figure 5).

In Figure 8, we show the 353 nodes of the Erdös set in the global K-K^*-plane (in a double logarithmic representation) with colored data points such that the color provides the value of $f_{rc}(K)$ for each node at given K-value (obtained at $n_{ib} = 1$ with the corresponding node as initial blue node). Optimal nodes with low red outcome (blue color) have typically small values in their K- and K^*-rank and nodes with large red outcome (red color) have typically large values in their K- and K^*-rank. Furthermore, Figure 8 also provides the positions of the 10 fixed red and the 6 blue nodes in the global K-K^*-plane with typically quite large values for K and K^*. Using the same data Appendix Figure A4 shows the dependence of f_{rc} (at $n_{ib} = 1$) on the rank K_{f_r}. Obviously, this curve is monotonically increasing and the range with small values of f_{rc} is rather small, i.e. $f_{rc} < 0.1$ corresponds to $K_{f_r} \leq 11$ (see also Table 2). Note that the average of the curve in Figure A4 corresponds to the data point at $n_{ib} = 1$ in Figure 4 which is $f_r(n_{ib} = 1) \approx 0.5818$.

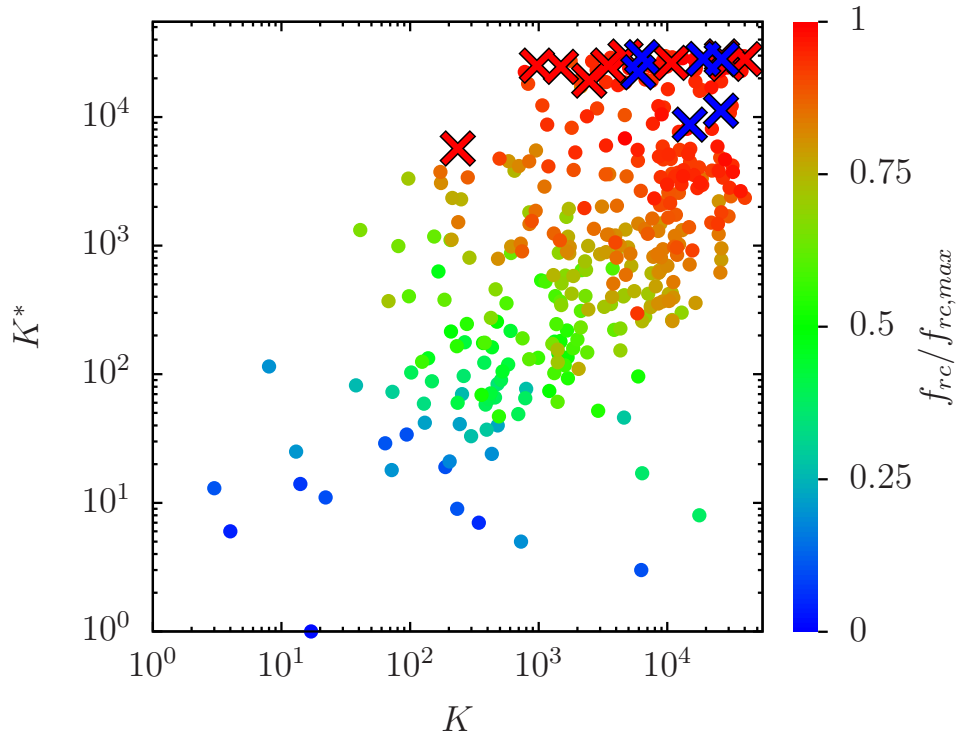


Figure 8. The positions of the 353 nodes of the Erös group in the K-K^* plane in a double logarithmic representation (colored full circles). The color of each data point corresponds to the value of $f_{rc}(K)$ using the initial configuration $[K]$ (i.e. $n_{ib} = 1$ with initial blue node at given value of K). To be more precise, the values of the color bar correspond to $f_{rc}(K)/f_{rc,max}$ with $f_{rc,max} \approx 0.77775$ being the maximum value of f_{rc} (i.e red for maximum, green for intermediate and blue for zero values and with no amplification of small values). The red (blue) crosses indicate the positions of the fixed 10 red (6 blue) nodes.

We have also analyzed the statistical properties of the individual $f_r(i)$ values, same quantity as in Figure 4, but using the data of Figure 6 for $n_{ib} = 4$ with the initial blue node configuration $[3, 4, 17, 342]$ and global (network averaged) value $f_r \approx 0.00318$. For example Appendix Figure A5 shows the 100 nodes with largest values of this quantity in the K-K^* plane. Obviously, among the 26 nodes with $f_r(i) = 1$ we have the 10 permanently fixed red nodes ($K_g = 1, \ldots, 10$ in Table 1) but there are also further 16 nodes with $f_r(i) = 1$ probably with (mostly) exclusive links to the fixed 10 red nodes, i.e. no or few links to other nodes, therefore explaining the fixed outcome $f_r(i) = 1$. There are about 1% (4%) of network nodes with values $f_r(i) > 0.5$ ($f_r(i) > 0.08850$).

Furthermore, Appendix Figure A6 shows the histogram distribution $p(f_r)$ for the same data with a rapid decay at $f_i < 1$ and a small peak at $f_r = 1$ corresponding to the 26 nodes i with $f_r(i) = 1$. We have also computed the related global probability $P(f_r)$ for a node i to have a value $f_r(i) > f_r$. This quantity is shown in Appendix Figure A7, confirming the rapid decay which is quite well algebraic as $P(f_r) \sim f_r^{-1.5}$ (for $f_r \geq 5 \times 10^{-3}$) and which is similar to the Poincare recurrences decay in symplectic chaotic maps with Ulam networks [31].

The slow algebraic decay of $P(f_r)$ has important consequences. It shows that the average values of f_r shown in Figure 6 even underestimate the effect. Explicitely, the average value of $f_r \approx 0.00318$ for $n_{ib} = 4$ in Figure 6 corresponds, according to Appendix Figure A7, to a fraction of nodes $P(0.00318) \approx 0.22$ having an $f_r(i)$ value above this average and 78% of nodes have a smaller $f_r(i)$ value. It is well known, that for such long tail distributions one should also focus on the median value $f_{r,median}$ defined by $P(f_{r,median}) = 0.5$ corresponding in our case to $f_{r,median} = 10^{-5}$ (only 1 red outcome in the $R = 100000$ pathway realizations) indicating that (slightly more than) 50% of nodes have an $f_r(i)$ value below or equal to the median value with nearly perfect reduction of the red outcome.

3.2. Results without formulas

In this work we presented a mathematical model of fibrosis progression in the PPI MetaCore network describing the global interaction structure of almost all proteins and important molecules (nodes). We show that even with only 10 fibrosis activated proteins the fibrosis progression can spread over a great majority of nodes (about 70 %). The developed analysis of network structure allows us to propose an efficient strategy which allows to reduce a number of fibrosis activated nodes by a factor 300 and disease elimination. The method is based on the Erdös barrage construction: we determine the Erdös nodes directly linked to the fixed 10 activated red nodes; this number can be relatively large (353 in our case); however, we show that a barrage with only 4 blue repairing Erdös nodes, corresponding to the four proteins c-Myc, p53, c-Fos and β-catenin, gives a reduction of the average number of fibrosis activated nodes by a factor 300; these 4 nodes belong to network nodes with high PageRank and CheiRank indexes of global MetaCore network.

Furthermore, this average actually underestimates the effect since it is determined by a relative small number of nodes with a modest reduction (e.g. factor of $\sim$ 100) while for more than 50% of nodes the reduction factor is even 100000 (only 1 infection outcome in the 100000 statistical realizations of our simulation). We have also identified in the group of Table 1 six interesting proteins HTR2B, ACSBG1, LGI2, ADORA2A, IL1R2-2 and COX4I2 (corresponding to the vertical green lines in Figures 5 and 7) where the reduction effect is somewhat less pronounced (compared to the 50% of nodes with nearly perfect reduction). We expect that our INFI model can be tested with other PPI networks (e.g. those of [32,33]).

4. Discussion and conclusion

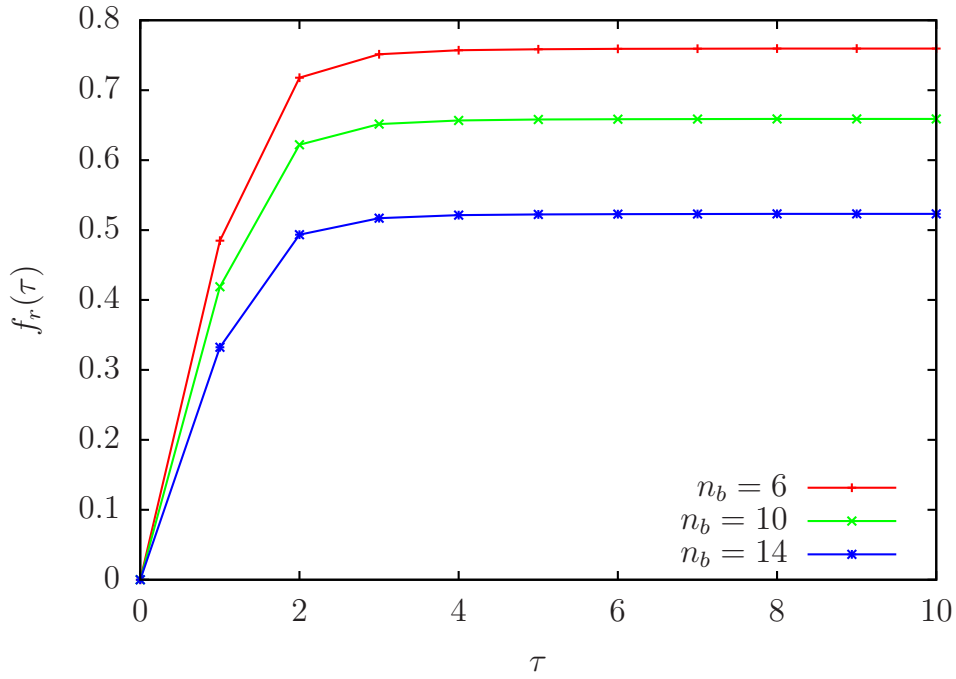
Identifying fibrosis-associated proteins and fibrosis progression is a critical issue in treating heart failure. An experimental determination of the fibrosis progression process is extremely time consuming and labor-intensive. In this work we present and describe a mathematical model of such a fibrosis progression using the global PPI MetaCore network. The performed analysis shows that even a small number of fibrosis activated proteins can lead to a global fibrosis progression of a major part of the whole PPI network.

We developed an efficient method of the Erdös barrage when a small group of e.g. 4 repairing protein can reduce the number of fibrosis activated proteins (nodes) by a factor 300 leading to a healthy state of the global system described here by the MetaCore network. We expect that similar results can be obtained for other disease progression. We think that it would be very interesting to test this INFI approach with other global PPI networks like [32,33]. We hope that the described INFI method will lead to new efficient medical treatments of fibrosis and other various diseases.

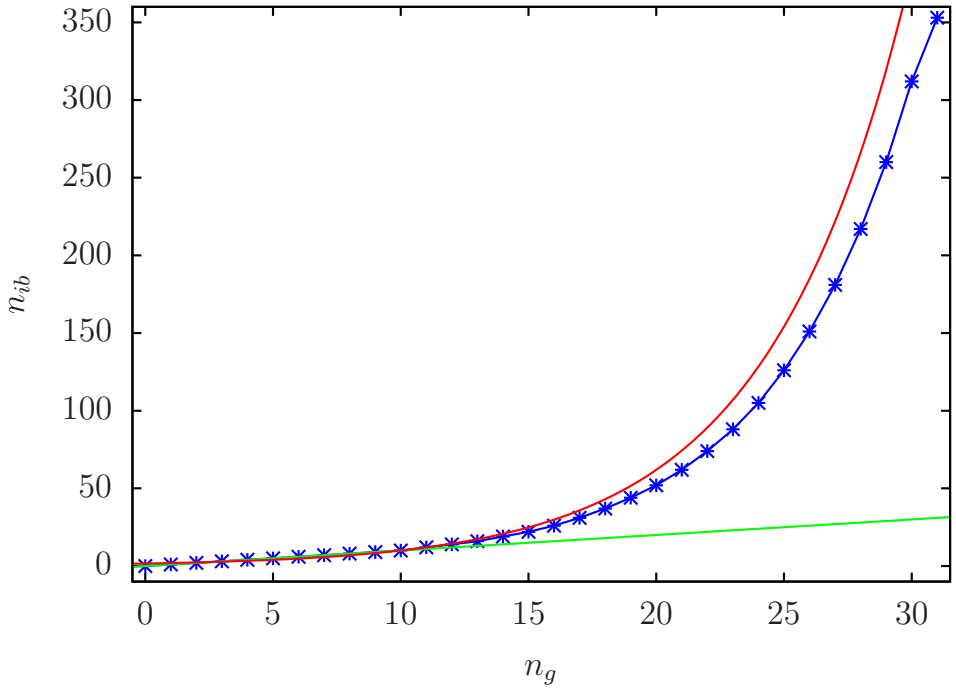
Appendix

A. Additional figures

Here we present additional Appendix Figures A1,. . . , A7 for the main part of this article.
Appendix Figure A1 illustrates the convergence of the time evolution of spin configurations.

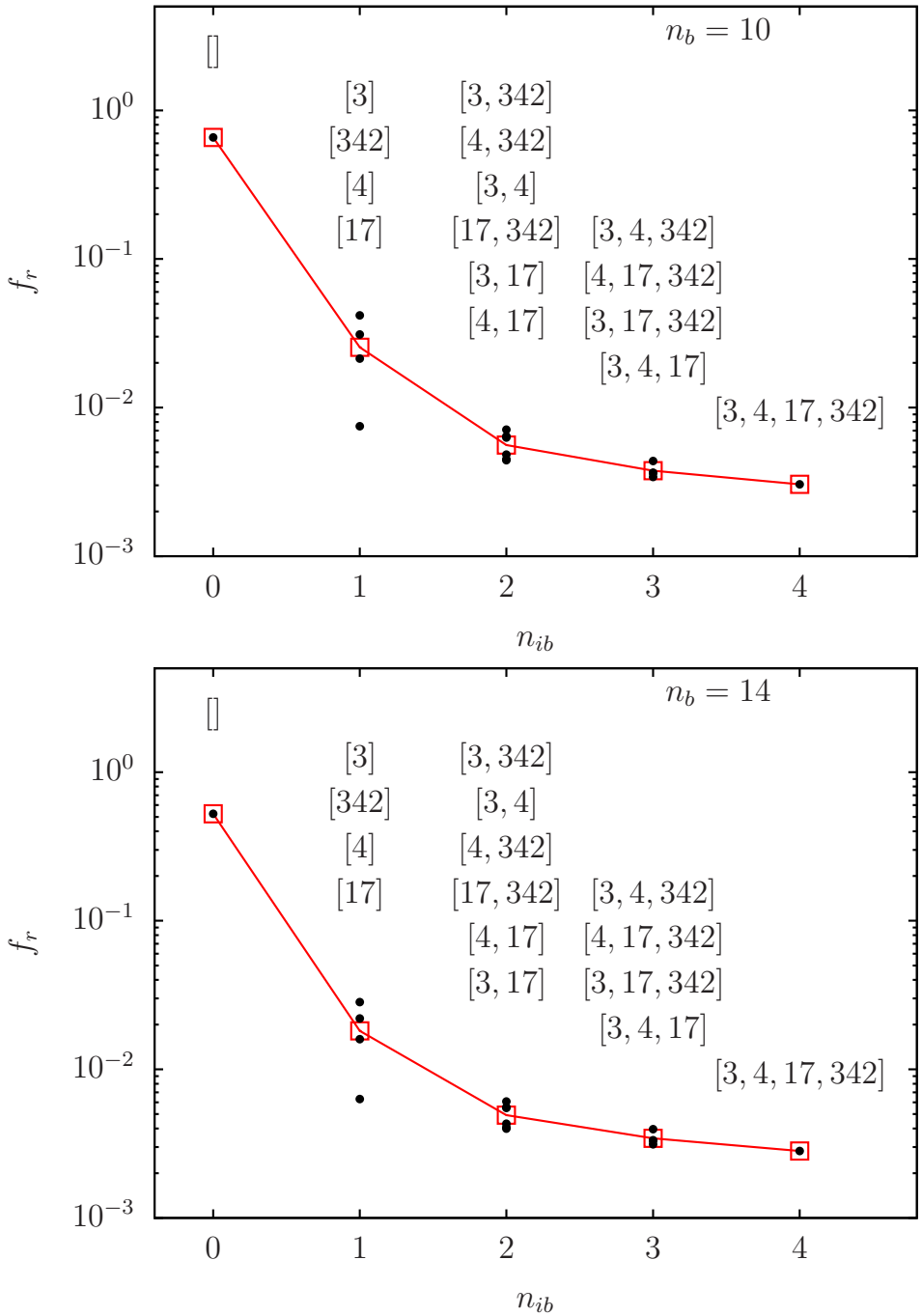


Appendix Figure A1. Illustration of the convergence of spin-configurations for $n_b = 6, 10, 14$ and $n_{ib} = 0$. Shown is average $f_r(\tau)$ versus iteration time τ. For practical reasons the convergence seems to be quite good at $\tau \approx 5$. However, the value of $f_r(\tau)$ becomes constant only at $\tau \geq \tau_{last} \approx 70$ with typical last non-vanishing differences $|f_r(\tau_{last} - 1) - f(\tau_{last})| \sim 10^{-10}$ and $|f_r(\tau_{last} + 1) - f(\tau_{last})| = 0$. All data in this work have been computed with values up to $\tau = 100$ to verify that exact convergence is achieved.



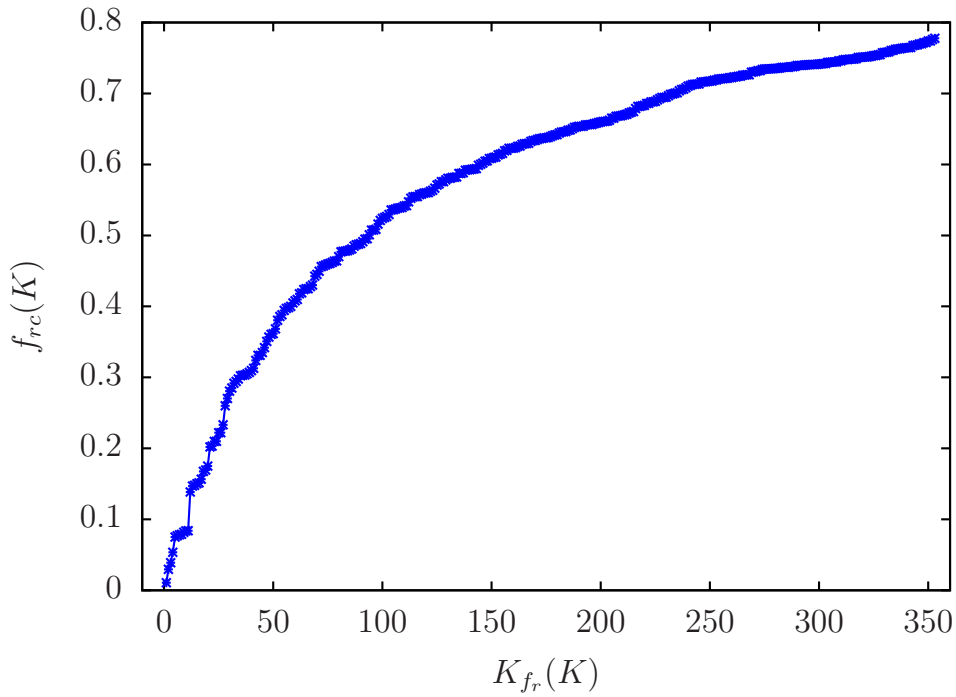
Appendix Figure A2. Dependence of n_{ib} on the coarse-grained index n_g used in Figure 5 (blue/star data points). The green line shows for comparison the linear behavior $n_{ib} = n_g$ (exactly identical to blue line for $n_g \leq 10$) and the red line shows the exponential behavior $n_{ib} = 10 \times 1.2^{n_g - 10}$ (good approximation for $n_g > 10$).

Appendix Figure A2 illustrates the link between n_{ib} and the coarse-grained index n_g used in Figure 5.



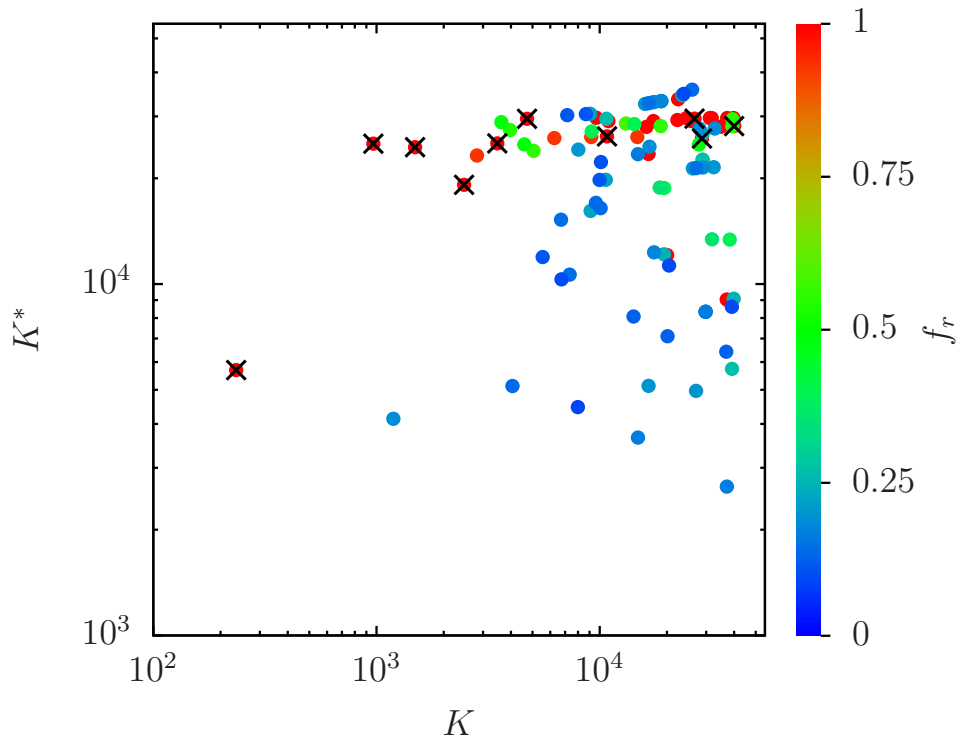
Appendix Figure A3. As Figure 6 but for $n_b = 10$ (top panel) and $n_b = 14$ (bottom panel).

Appendix Figure A3 shows f_r for optimal configurations at $n_b = 10$ and $n_b = 14$ (as Figure 6).



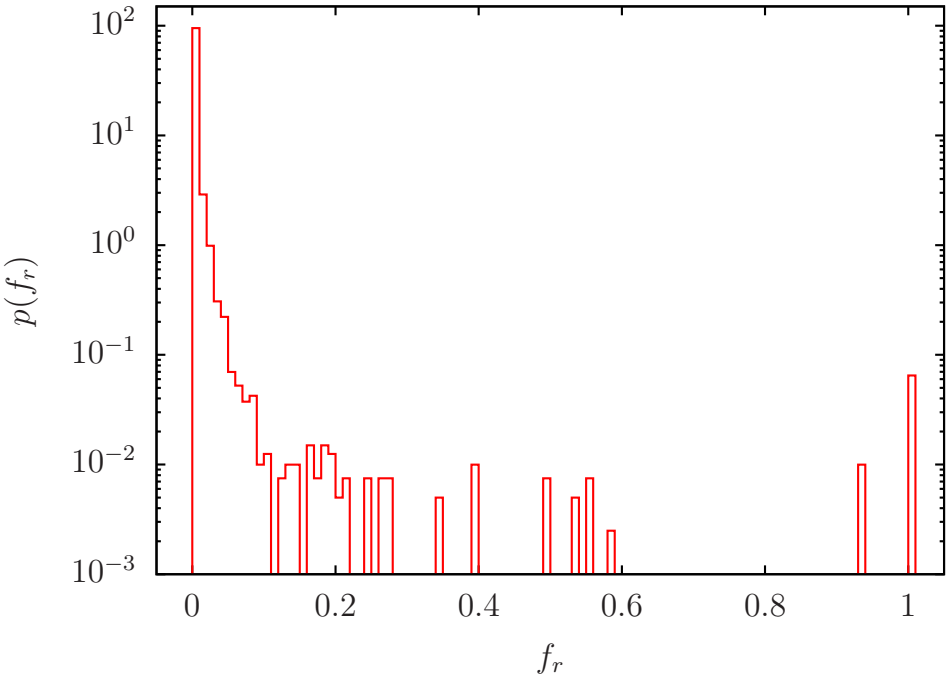
Appendix Figure A4. $f_{rc}(K)$ versus $K_{f_r}(K)$ using the data of Figure 8 (i.e. $n_b = 6$, $f_{rc}(K)$ being computed for $n_{ib} = 1$ with one single initial blue node at node K belonging to the Erdös group.) The index $K_{f_r}(K)$ is the index obtained by ordering the $N_E = 353$ nodes of the Erdös set with increasing values of $f_{rc}(K)$.

Appendix Figure A4 shows the values of $f_{rc}(K)$ (for $n_b = 6$) computed from one initial (variable) blue node K belonging to the Erdös group ordered by increasing values of $f_{rc}(K)$ inside this group.



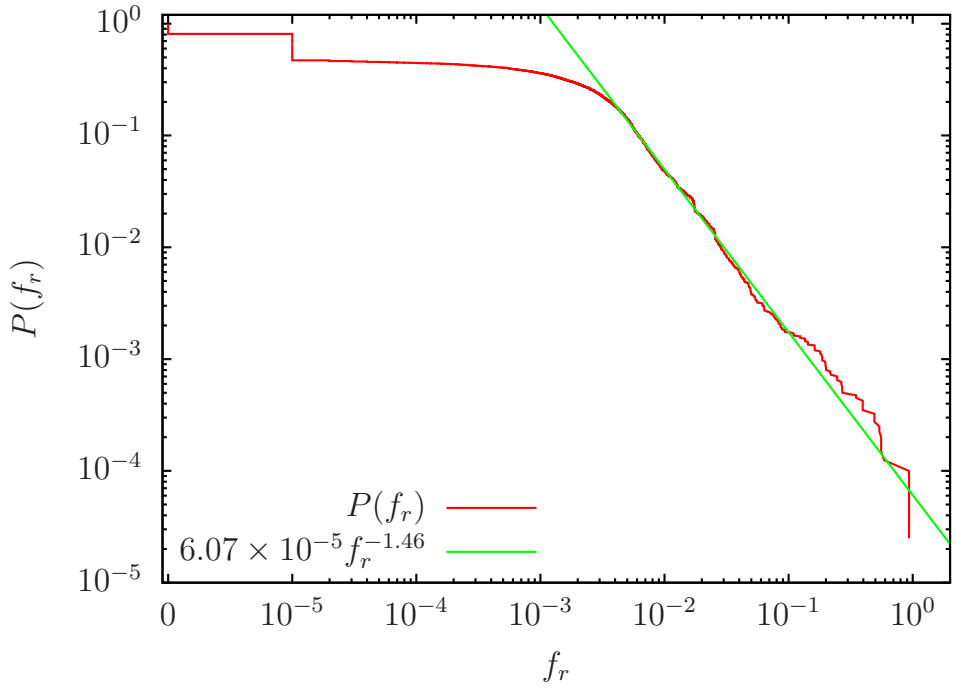
Appendix Figure A5. The positions of the 100 nodes with top 100 $f_r(i)$ values of the (non network averaged) red outcome for individual nodes i in the K-K^* plane in a zoomed double logarithmic representation (colored full circles) using the data for $n_{ib} = 4$ of Figure 6 with the configuration $[3, 4, 17, 342]$ for the initial blue nodes. The color of each data point corresponds to the value of $f_r(i)$ according to the color bar. There are 26 nodes i with $f_r(i) = 1$ including the 10 permanently fixed red nodes (black cross symbols; $K_g = 1, \ldots, 10$ in Table 1) and 16 additional nodes and there are 36 (100) nodes i with $f_r(i) > 0.5$ ($f_r(i) \geq 0.08850$).

Appendix Figure A5 shows for the data for $n_{ib} = 4$ of Figure 6 with the initial blue node configuration $[3, 4, 7, 342]$ the 100 nodes with top individual $f_r(i)$ values of the red outcome in the K-K^* plane.



Appendix Figure A6. Histogram distribution $p(f_r)$ for the data of Figure A5 using the values $f_r(i)$ for all nodes with bin width $\Delta f_r = 0.01$. The normalization is as in Figure 1 by an integral.

Appendix Figure A6 shows the distribution $p(f_r)$ for the data of Figure A5. The peak at $f_r = 1$ corresponds to the 26 nodes i with $f_r(i) = 1$. The large peak at $f_r = 0$ corresponds to 95% of probability for $0 \leq f_r \leq 0.01$.



Appendix Figure A7. Fraction $P(f_r)$ of nodes i with $f_r(i) \geq f_r$ versus f_r (red curve) for the data of Figs. A5, A6. The presentation is double logarithmic except for the data points at $f_r = 0$ which have artificially been added at a finite value below 10^{-5} and the 26 nodes i with $f_r(i) = 1$ have been excluded in the computation of $P(f_r)$. The steps/vertical lines at $f_i = 0$ ($f_i = 10^{-5}$) correspond to 7539 (13696) network nodes i having 0 (1) case(s) of red outcome out of the $R = 100000$ random pathway realizations (note that there are 7230 nodes who have always white outcome). The green straight line corresponds to the power law fit $P(f_r) \approx a f_r^b$ with $a = (6.07 \pm 0.03) \times 10^{-5}$ and $b = -1.457 \pm 0.001$ obtained for the interval $f_i \in [5 \times 10^{-3}, 4 \times 10^{-2}]$.

Appendix Figure A7 shows the probability $P(f_r)$ for a node i to have a value $f_r(r) \geq f_r$ for the data of Figs. A5, A6. The decay of $P(f_r)$ is rather well algebraic as $P(f_r) \approx 6.07 \times 10^{-5} f_r^{-1.46}$ for $f_r \geq 5 \times 10^{-3}$ if the 26 nodes i with $f_r(i) = 1$ are excluded in the computation of $P(f_r)$. (If these nodes are included the decay is still algebraic in the fit interval $f_i \in [5 \times 10^{-3}, 4 \times 10^{-2}]$ but $P(f_r)$ is significantly above the power law for $f_i > 4 \times 10^{-2}$.) Note that $P(f_r)$ corresponds to the integrated probability $P(f_r) = \int_{f_r}^1 p(\tilde{f}_r)\, d\tilde{f}_r$ using the probability density $p(f_r)$ visible in Appendix Figure A6 (apart from the numerical approximation due to the histogram with finite bin width).

Author Contributions: All authors equally contributed to all stages of this work.

Funding: This research was supported in part through the grant NANOX N^o ANR-17-EURE-0009, (project MTDINA) in the frame of the Programme des Investissements d'Avenir, France; it was granted access to the HPC resources of CALMIP (Toulouse) under the allocation 2024-P0110; it was also supported by INSERM funding.

Acknowledgments: We thank L.Ermann for useful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Murtha L.A., Schuliga M.J., Mabotuwana N.S, Hardy S.A., Waters D.W., Burgess J.K., Knight D.A. and Boyle A.J., *The processes and mechanisms of cardiac and pulmonary fibrosis*, Front Physiol. **128** (2017) 777.
2. Liu T., Song D., Dong J., Zhu P., Liu J., Liu W., Ma X., Zhao L. and Ling S., *Current understanding of the pathophysiology of myocardial fibrosis and its quantitative assessment in heart failure*, Front. Physiol. **8** (2017) 238.
3. Meng Xm., Nikolic-Paterson D. and Lan H. *TGF-β: the master regulator of fibrosis*, Nat. Rev. Nephrol. **12** (2016) 325.
4. Wynn T.A., *Cellular and molecular mechanisms of fibrosis*, Jour. Pathology **214(2)** (2007) 199.

5. Pintus S.S., Sharipov R.N., Kel A., Timotin A., Keita S., Martinelli I., Boal F., Tronchere H., Kolpakov F. and Kunduzova O., *Drug repositioning for cardiac fibrosis through molecular signature of aberrant fibroblast activation*, INSERM preprint, to be published (2021).
6. Karimizadeh E., Sharifi-Zarchi A., Nikaein H., Salehi S., Salamatian B., Elmi N., Gharibdoost F. and Mahmoudi M., *Analysis of gene expression profiles and protein-protein interaction networks in multiple tissues of systemic sclerosis*, BMC Medical Genomics **12** (2019) 199.
7. Pchejetski D., Foussal C., Alfarano C., Lairez O., Calise D., Guilbeau-Frugier C., Schaak S., Seguelas M-H., Wanecq E., Valet P., Parini A. and Kunduzova O., *Apelin prevents cardiac fibroblast activation and collagen production through inhibition of sphingosine kinase 1*, Eur. Heart J. **33** (2012) 2360
8. MetaCore, Available: https://clarivate.com/cortellis/solutions/early-research-intelligence-solutions/. Accessed September 2024.
9. Ekins S., Bugrim A., Brovold L., Kirillov E., Nikolsky Y., Rakhmatulin E., Sorokina S., Ryabov A., Serebryiskaya T., Melnikov A., Metz J. and Nikolskaya T., *Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms*, Xenobiotica **36(10-11)** (2006) 877; doi:10.1080/00498250600861660.
10. Bessarabova M., Ishkin A., JeBailey L., Nikolskaya T., and Nikolsky Y., *Knowledge-based analysis of proteomics data*, BMC bioinformatics **13 Suppl 16** (2012) 13; doi:10.1186/1471-2105-13-S16-S13.
11. Kotelnikova E., Frahm K.M., Lages J. and Shepelyansky D.L., *Statistical properties of the MetaCore network of protein-protein interactions*, Appl. Netw. Sci **7** (2022) 7.
12. Kotelnikova E., Frahm K.M., Shepelyansky D.L. and Kunduzova O., *Fibrosis protein-protein interactions from Google matrix analysis of MetaCore network*, Int. J. Mol. Sci. **23** (2022) 67.
13. Brin S. and Page L., *The anatomy of a large-scale hypertextual Web search engine*, Computer Networks and ISDN Systems **30** (1998) 107.
14. Langville A.M. and Meyer C.D., *Google's PageRank and beyond: the science of search engine rankings*, Princeton University Press, Princeton (2006).
15. Ermann L., Frahm K.M. and Shepelyansky D.L., *Google matrix analysis of directed networks*, Rev. Mod. Phys. **87** (2015) 1261.
16. Dorogovtsev, S. *Lectures in Complex Networks*; Oxford University Press, 2010.
17. Castellano, C.; Fortunato, S.; Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys.* **2009**, *81*, 591–646. https://doi.org/10.1103/RevModPhys.81.591.
18. Galam, S. Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach. *Journal of Mathematical Psychology* **1986**, *30*, 426–434. https://doi.org/https://doi.org/10.1016/0022-2496(86)90019-2.
19. Sznajd-Weron, K.; Sznajd, J. Opinion evolution in closed community. *International Journal of Modern Physics C* **2000**, *11*, 1157–1165, [https://doi.org/10.1142/S0129183100000936]. https://doi.org/10.1142/S0129183100000936.
20. Sood, V.; Redner, S. Voter Model on Heterogeneous Graphs. *Phys. Rev. Lett.* **2005**, *94*, 178701. https://doi.org/10.1103/PhysRevLett.94.178701.
21. Watts, D.J.; Dodds, P.S. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research* **2007**, *34*, 441–458, [https://doi.org/10.1086/518527]. https://doi.org/10.1086/518527.
22. Galam, S. Sociophysics: a review of Galam models. *International Journal of Modern Physics C* **2008**, *19*, 409–440, [https://doi.org/10.1142/S0129183108012297]. https://doi.org/10.1142/S0129183108012297.
23. Schmittmann, B.; Mukhopadhyay, A. Opinion formation on adaptive networks with intensive average degree. *Phys. Rev. E* **2010**, *82*, 066104. https://doi.org/10.1103/PhysRevE.82.066104.
24. Kandiah, V.; Shepelyansky, D.L. PageRank model of opinion formation on social networks. *Physica A* **2012**, *391*, 5779. https://doi.org/10.1016/j.physa.2012.06.047.
25. Coquide, C.; Lages, J.; Shepelyansky, D.L. Opinion Formation in the World Trade Network. *Entropy* **2024**, *25(2)*, 141. https://doi.org/10.3390/e26020141.
26. Ermann, L.; Shepelyansky, D.L. Confrontation of capitalism and socialism in Wikipedia Networks. *arXiv:2408.07606[cs.SI]* **2024** https://doi.org/10.48550/arXiv.2408.07606.
27. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci.* **1982**, *79(8)*, 2554. https://doi.org/10.1073/pnas.79.8.2554.
28. Benedetti, M.; Carillo, L.; Marinari, E.; Mezard, N. Eigenvector dreaming. *J. Stat. Mech.* **2024**, 013302. https://doi.org/10.1088/1742-5468/ad138e.

29. Albert, R.; Thakar, J. Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *WIREs Syst. Biol. Med.* **2014**, *6*, 353. https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.1273.
30. Tripathi, S.; Kessler, D.A.; Levine, H. Biological Networks Regulating Cell Fate Choice Are Minimally Frustrated. *Phys. Rev. Lett.* **2020**, *125*, 088101. https://doi.org/10.1088/1742-5468/ad138e.
31. Frahm, K.M.; Shepelyansky, D.L. Poincare recurrences and Ulam method for the Chirikov standard map. *Eur. Phys. J. B* **2013**, *86*, 322. https://link.springer.com/article/10.1140/epjb/e2013-40120-6.
32. TRANSPATH, Available: https://genexplain.com/transpath/. Accessed September 2024.
33. REACTOME, Available: https://reactome.org/. Accessed September 2024.