Fast randomisation of large genomic datasets while preserving alteration counts

Andrea Gobbi[1,5,*], Francesco Iorio[2,3,*,†], Kevin J. Dawson[3], David C. Wedge[3], David Tamborero[4], Ludmil B. Alexandrov[3], Nuria Lopez-Bigas[4], Mathew J. Garnett[3], Giuseppe Jurman[1], Julio Saez-Rodriguez[2]

[1]Fondazione Bruno Kessler, Trento – Italy, [2]European Molecular Biology Laboratory – European Bioinformatics Institute, Cambridge – UK, [3]Wellcome Trust Sanger Institute, Cambridge – UK, [4]Universitat Pompeu Fabra, Barcelona – Spain, [5]University of Trento, Trento – Italy.

Supplementary Information

1.1 Efficient implementation of the Switching-Algorithm

Let $G = (V, E)$ a bipartite graph with node sets V_r and V_c.

Let B be the incidence matrix of G, i.e. a $V_r \times V_c$ binary matrix such that $w_{i,j} = 1$ if and only if the i-th node of the first class is connected to the j-th node of the second class.

Finally, let L be the edge-list of B, i.e. $|E| \times 2$ matrix such that $(L_{i,1}, L_{i,2}) \in E \ \ \forall\, i = 1, \dots, |E|$.

Let us suppose that we want to perform N switching-steps then, our optimal implementation of the switching-algorithm proceeds as follows:

Algorithm 4.1: switching-algorithm(G, N)

$f \leftarrow edgeList(G)$

$B \leftarrow incidenceMatrix(G)$

repeat

Randomly choose $m, n \in [1, \dots, |E|],\ m \neq n$

$a \leftarrow L_{n,1}, b \leftarrow L_{n,2}$

$c \leftarrow L_{m,1}, d \leftarrow L_{m,2}$

if $w_{a,d} = 0, w_{c,b} = 0, b \neq c, a \neq d, a \neq b, c \neq d$

then $\begin{cases} w_{a,d} \leftarrow 1,\ w_{c,b} \leftarrow 1 \\ w_{a,b} \leftarrow 0,\ w_{c,d} \leftarrow 0 \\ L_{n,1} = d,\ L_{m,2} = b \end{cases}$

$N \leftarrow N - 1$

until $N = 0$

$G' \leftarrow graph(B)$

return (G')

1.2 Expected similarity between any pair of random bipartite networks with prescribed number of nodes and edges but with possibly different node degrees

The number of common edges between two random bipartite networks $G_1 = (\{V_r, V_c\}, E_1)$, and $G_2 = (\{V_r, V_c\}, E_2)$, containing both $|E_1| = |E_2| = |E|$ edges, follows the hypergeometric distribution

$$P\left(\left|E_2 \bigcap E_1\right| = x\right) = \frac{\binom{|E|}{x}\binom{t-|E|}{|E|-x}}{\binom{t}{|E|}}$$

(where $t = |V_r| \times |V_c|$ is the number of possible edges preserving bipartiteness), whose mean value is equal to $|E|^2/t$. Thus replacing this value in equation 3 of the main text, the average similarity between the original network G and one of its fully-randomised versions is equal to:

$$s = \frac{|E|}{2t - |E|}$$

1.3 Convergence criteria and autocorrelation time

The autocorrelation time is a quantity related to the mixing-time of a Markov chain (Sokal, 1989). Briefly, the autocorrelation of a signal is the cross-correlation of that signal with itself given a lag T. Formally, for a series of data $\langle X_i \rangle$, where each X_i is a drawn from the same distribution with mean μ and variance σ, the autocorrelation is given by $R_X(T) = E[(X_i - \mu)(X_{i-T} - \mu)]/\sigma^2$ (here $E(.)$ indicates the mean function).

When drawing independent samples from the stationary distribution of the Markov chain underlying a sampler, the autocorrelation of that set of samples with itself would tend to 0 as the number of samples increases (Stanton and Pinar, 2012). As a consequence, the autocorrelation time (i.e. the lag needed for the autocorrelation to reach its minimum) captures the size of the gaps between sampled states of the chain needed before the autocorrelation of this "thinned" chain is very small. If the chain has 0 autocorrelation then it is sampling from its stationary distribution. In (Stanton and Pinar, 2012), the authors uses the autocorrelation time as estimation of the mixing time because they measure the same thing: the number of iterations needed by the Markov chain in order for the difference between its current distribution and the stationary distribution to be small.

To show that after N switching-step the average edge autocorrelation of the tested networks (as defined in (Stanton and Pinar, 2012)) fluctuates around its minimal value after an exponential drop off, we ran different instances of the switching-algorithm on the simulated networks described in section 3.1 (fixed node set sizes and variable edge-density). For each of these networks, the switching-algorithm was executed for a total number of $50N$ switching-steps. Individual binary signals were derived for a set of randomly selected edges (whose cardinality was equal to 10% of the total number of existing edges in the original network under consideration). Each of these signals was composed of the entries corresponding to the edge under consideration in the B^k BEMs, for each $k = 1, \ldots, 10N$. Finally, each signal X was sampled with different lags $T = 50i$, with $i = 1, \ldots, \lceil 10N/50 \rceil$ and the autocorrelation $R_X(T)$ was computed within each lag.

Results of these simulations are shown in supplementary figure S2 (left side). For all the three networks increasing the lag time causes an exponential drop in the average edge autocorrelation, which then fluctuates around zero. In all the three cases (corresponding to networks with different edge densities, indicated by different colours), the

stochastic behavior around zero starts before the lag time equal to N (dashed line in the plots and equal to 20,177 for an edge density of 20%, 45,353 for 35% and 42,585 for 50%).

Additionally, as shown in figure 3 of the main text (right side) we observed an almost perfect correlation (> 0.99, for all the three cases) between the average edge autocorrelation $R_X(T)$ and the average Jaccard index (JI, as defined in the previous sections) computed between each pair of BEMs composed by $B^{k(T-1)}$ and its rewired version after T switching-steps B^{kT}, for each T (with B^0 equal to the original starting BEM).

Additionally, as shown in the inset of supplementary figure S2 (and the points plotted in red) for lag times greater or equal than N, both the average edge autocorrelation and the average Jaccard index fluctuate around their minimum.

These results suggest that our convergence criteria can be considered as a good estimator for autocorrelation time, and hence the mixing time, of the Markov chain underlying the switching-algorithm.

1.4 Comparison with the empirical bound

We conducted an empirical study to show that after N switching steps the initial bias of the Markov chain underlying the switching-algorithm, quantified by the residual similarity to the original network (i.e. $x^{(k)}$), is minimised at least as much as it is minimised after $N'=100e$ switching steps (i.e. the empirical bound proposed by (Milo *et al.*, 2003)). Specifically, we executed 2,500 independent runs of the switching-algorithm on an incidence matrix modeling a bi-partite network with $n_c = 500$, $n_r = 1{,}000$, and an edge density d equal to ~4%.

We considered as a reference 'stationary distribution' of the $x^{(k)}$ values the one reached after $N' = 100e$ switching steps.

Results of this simulation are depicted in supplementary figure S3 (A): here the black curve indicates the difference between the number of edges shared by the original network and its rewired versions at the k-th and the $(k-1)$-th switching steps, respectively $x^{(k)}$ and $x^{(k-1)}$, averaged across the 2,500 independent runs of the switching-algorithm. Consistently with the simulations presented in the previous sections, this difference reaches a plateau, close to zero, before N switching steps (in this case equal to 102,285) and far more before N' switching steps (in this case equal to 1,992,300). The same happens to the Total variation distance and the Kolmogorov distance (respectively blue and red curves in supplementary figure S3 (A)) between the distribution of the $x^{(k)}$ values and that of the $x^{(N')}$ values across the 2,500 independent runs of the switching-algorithm. The trend of the first 5 moments of the $x^{(k)}$ values distribution, in function of the number of switching steps, confirming a convergence time lower than N and far more lower than N', is provided in supplementary figure S4.

In supplementary figure S3 (B) the evolution of the distribution of edges in common between the current rewired version of the network and the original one is shown. The color code reflect the number of performed switching step and in black is depicted the limit distribution reached at N', which is equal to that reached at N.

1.5 Breast cancer dataset pre-processing and binary event matrix construction

Breast cancer samples and their respective mutations were downloaded from the Cancer Genome Atlas (TCGA) projects data portal (http://tcga.cancer.gov/dataportal/).

Germline mutations were filtered out from this dataset using a subset of germline mutations from dbSNP (Sherry, 2001), the 1000 genomes project (1000 Genomes Project Consortium *et al.*, 2012), and the NHLBI GO Exome Sequencing Project (Fu *et al.*, 2012). Optimised cutoffs of minor allele frequencies were applied for pre-filtering the three germline datasets in order to avoid removing of any known driver mutations. Further, a somatic mutation

at the same genomic position as a germline variant was removed only if it was exactly matched the type of germline variant. Synonymous mutations and mutations identified as benign and tolerated, respectively, by SIFT (Kumar *et al.*, 2009) and PolyPhen (Adzhubei *et al.*, 2010) were also removed from the dataset.

A binary event matrix (supplementary data DS1) was constructed from the remaining deleterious somatic mutations, yielding 757 rows (i.e. samples), 9,757 columns (i.e. genes), 19,758 non-null entries (i.e. variants), corresponding to an edge density equal to 0.27% in the corresponding bipartite network.

1.6 Colorectal cancer dataset pre-processing and binary event matrix construction

We analysed mutual exclusivity patterns for the protein affecting mutations of a colorectal cancer dataset assembled from the TCGA and other studies, by using the consequence type retrieved from the Ensembl variant effect predictor tool (Chen et al., 2010). We limited this analysis to those genes identified as putative mutational drivers by following a similar approach to that described in (Tamborero, Gonzalez-Perez, Perez-Llamas, et al., 2013) in which several methods aiming at detecting complementary signals of positive selection were combined.

We used MutSigCV (Lawrence et al., 2013) to identify recurrently mutated genes, OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012) to detect genes biased towards the accumulation of mutations with a larger functional impact, and OncodriveCLUST (Tamborero, Gonzalez-Perez, and Lopez-Bigas, 2013) to select genes exhibiting larger than expected by random chance mutation accumulations across the protein sequence. Finally we removed genes altered in less than 2 samples.

1.7 Time complexity analysis of the *Rewire* function contained in the package *iGraph* v0.6.1

The *rewire* function contained in *igraph* v0.6.1 does not implement the switching-algorithm but proceeds through a series of rewiring steps as follows:

Given a bipartite network $G = (V, E)$ with node sets V_r and V_c,

1. two nodes a, c are randomly selected from V_r;
2. b is randomly selected among the nodes in V_c that are connected to a (i.e. neighbours of a);
3. d is randomly selected among the nodes in V_c that are connected to c (i.e. neighbours of c).

This strategy systematically biases the edge selection, by privileging edges connected to nodes in V_c that have low degree. In fact, in the first step of the above list, each node has the same probability of being selected but in step 2 and 3, the probability of extracting a neighbour of a selected node (hence a given edge) is inversely proportional to the degree of that node. Specifically, with this strategy the distribution of selected edges is uniform only if the degree distribution of the nodes in V_c is uniform (i.e. each node has the same number of incident edges).

Additionally, this implementation requires, at each step, a local exploration of the network that is generally slower than storing and retrieving individual edges from an edge list. In particular, the asymptotic time complexity for a single switching step in the first case would be $O(|V| + |E|)$ while it would be proportional to the maximal observed degree (i.e. $O(\max(degree)) = O(|V|)$ in the second case. As a consequence, performing N of these steps does not guarantee that the residual similarity reaches its minimal value (as shown in the second column of Table 1 (B) of the main text) and the execution time of our implementation of the switching-algorithm is significantly lower than the one required by the *rewire* function (i.e. $\sim 3.2 \times 10^3$ sec vs. $\sim 1.8 \times 10^4$ sec to execute N steps, $\sim 3.46 \times 10^4$ sec vs. $\sim 3.73 \times 10^5$ sec to execute N' steps).

Supplementary datasets DS1 and DS2 are available on the BiRewire website:

http://www.ebi.ac.uk/~iorio/BiRewire/BiRewire/BiRewire_Home.html

at:

http://www.ebi.ac.uk/~iorio/BiRewire/BiRewire/BiRewire_Home_files/SuppData_SD1_BRCA_dataset.txt

and

http://www.ebi.ac.uk/~iorio/BiRewire/BiRewire/BiRewire_Home_files/SuppData_SD2_COREAD_dataset.txt

References:

1000 Genomes Project Consortium *et al*. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Adzhubei,I.A. *et al*. (2010) A method and server for predicting damaging missense mutations. *Nature Methods*, **7**, 248–249.

Chen,Y. et al. (2010) BMC Genomics | Full text | Ensembl variation resources. BMC ….

Fu,W. *et al*. (2012) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.

Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. Nucleic Acids Res, 40, e169.

Kumar,P. *et al*. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, **4**, 1073–1081.

Lawrence,M.S. et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499, 214–218.

Milo,R. *et al*. (2003) On the uniform generation of random graphs with prescribed degree sequences. *Arxiv preprint cond-mat/0312028*.

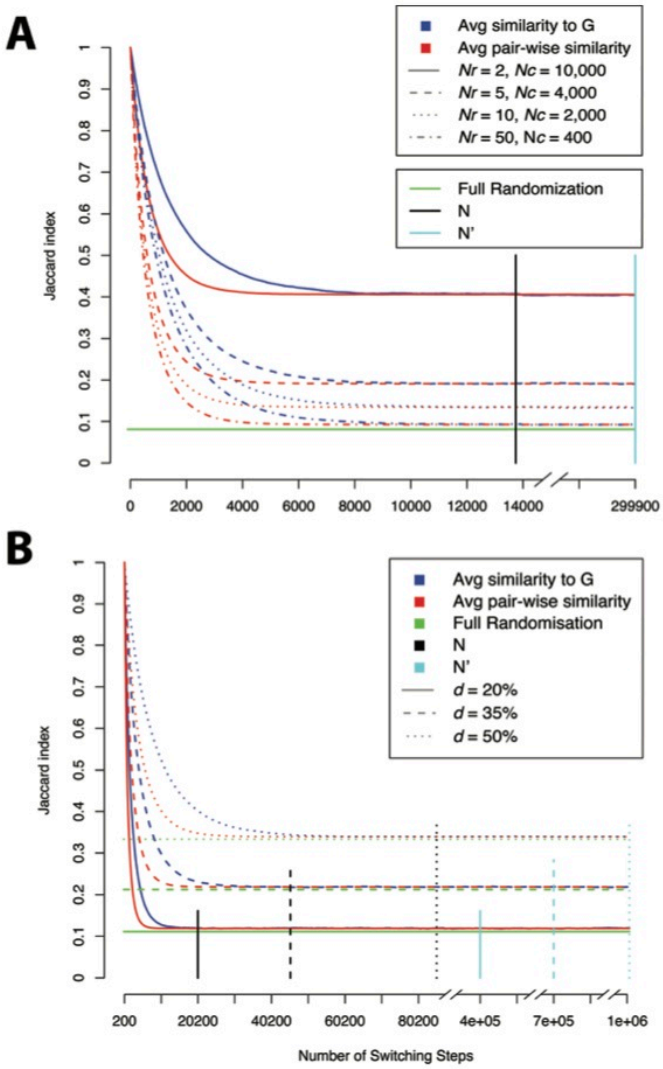
Sherry,S.T. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308–311.

Sokal,A.D. (1989) Monte Carlo methods in statistical mechanics: foundations and new algorithms.

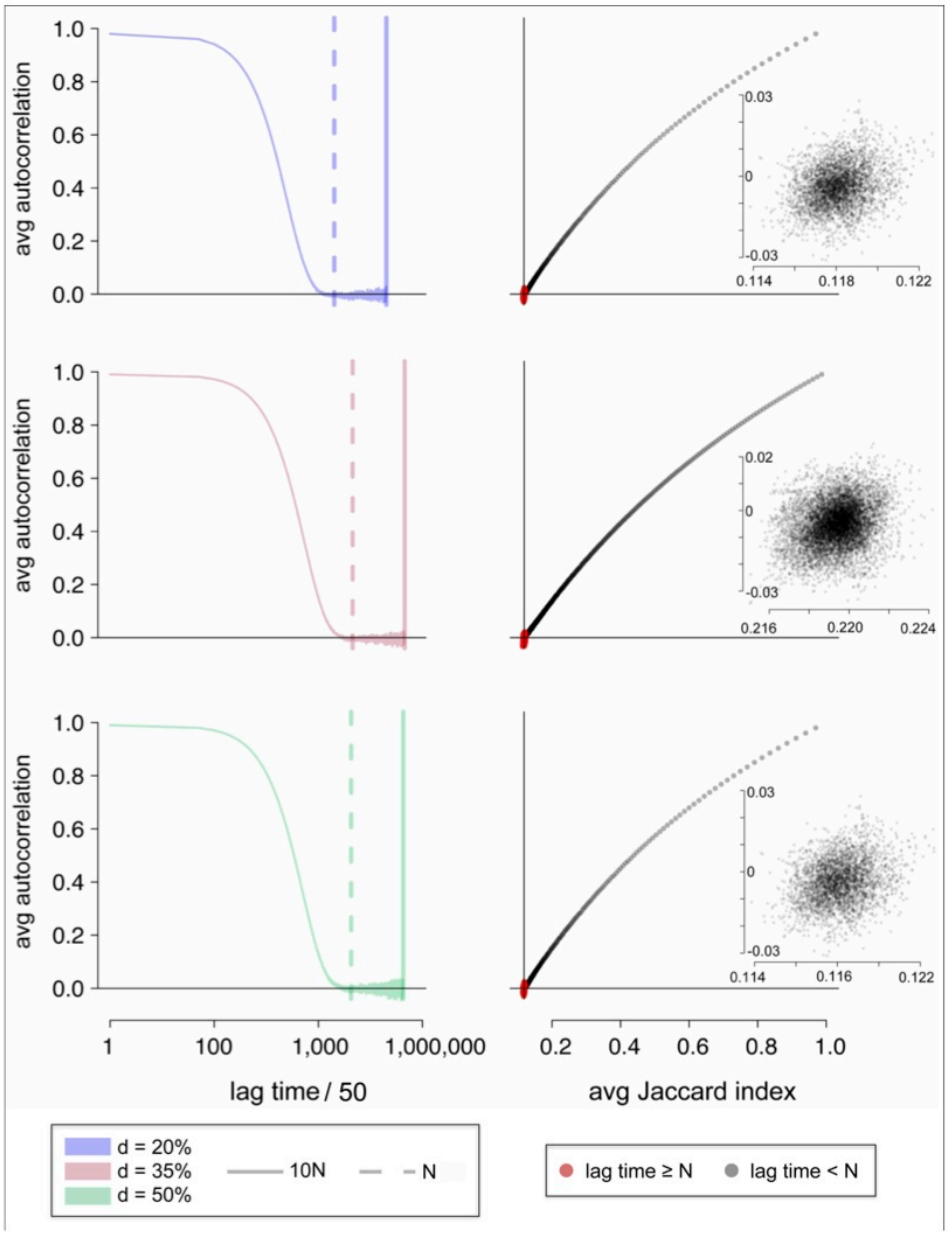
Stanton,I. and Pinar,A. (2012) Constructing and sampling graphs with a prescribed joint degree distribution. *J. Exp. Algorithmics*, **17**, 3.1.

Tamborero,D., Gonzalez-Perez,A., and Lopez-Bigas,N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics, 29, 2238–2244.

Tamborero,D., Gonzalez-Perez,A., Perez-Llamas,C., et al. (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep, 3, 2650.



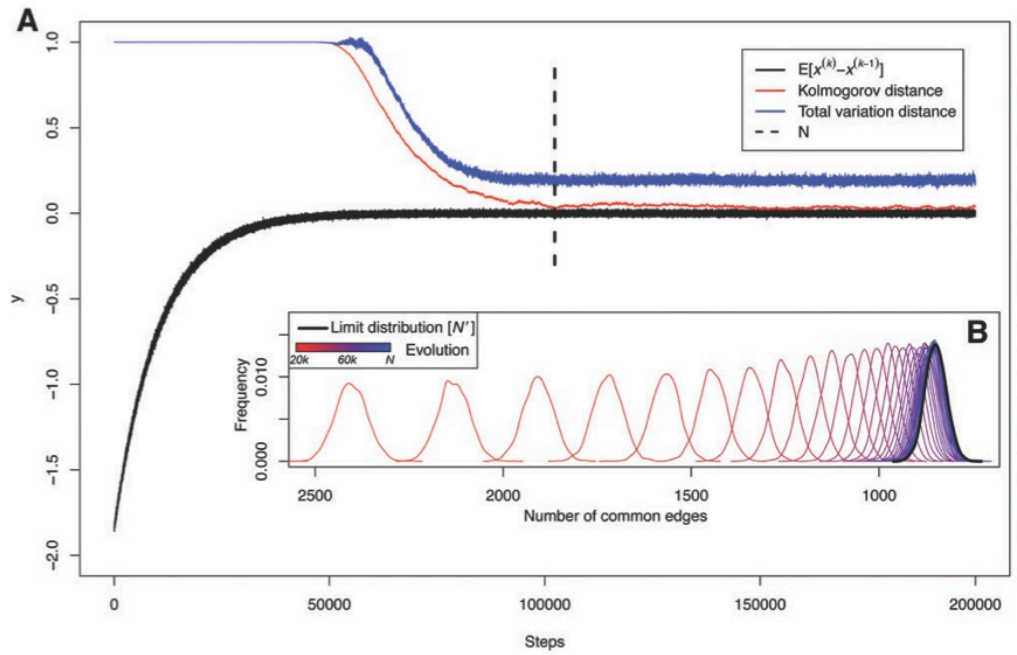
Supplementary Figure S1 - Trends of similarity to the original network as function of the number of switching-steps. The vertical lines indicate the number of switching-steps to be performed according to our new lower bound (black line) and the empirical one (cyan line). Each blue curve indicates the average similarity between the original bipartite network under consideration and its 50 different rewired versions at different sample time (i.e. increasing numbers of switching-steps). Each red curve indicates the average pair-wise similarity between each pair of rewired networks. The green curves indicate the expected similarity between any two random bipartite networks with the same numbers of nodes, density and squareness of the networks under consideration. (A) Different line styles refer to bipartite networks with different level of squareness. All the original networks contained 20,000 nodes and 3,000 edges. (B) Different line styles refer to bipartite networks with different levels of edge density. All the original networks contained two classes of 100 and 200 nodes, respectively.



Supplementary Figure S2 – Comparison between autocorrelation and our convergence criterion.

Plots on the left show the average auto-correlation as function of the lag time for 3 networks with different edge density (indicated by the colors). For all the three networks increasing the lag time causes an exponential drop in the average edge autocorrelation, which then fluctuates around zero. In all the three cases the stochastic behavior around zero starts before the lag time equal to *N* (dashed line in the plots). On the right scatter plots of the average Jaccard index and the average edge autocorrelation computed between each pair of BEM and its rewired version at T switching-steps/lag-time show an almost perfect correlation between these two convergence diagnostic metrics.

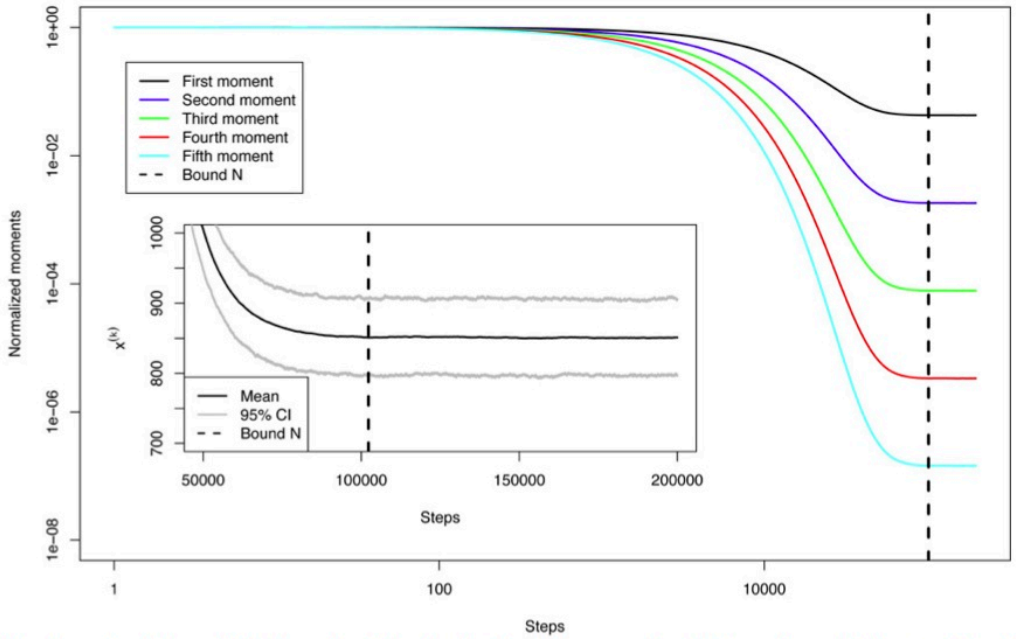
In the inset of the right plots, is a magnification of the region containing the points in red (i.e. values of the metrics for lag times greater or equal than N): both the average edge autocorrelation and the average Jaccard index fluctuate around their minimum.



Supplementary Figure S3 – Network similarity distributions comparison

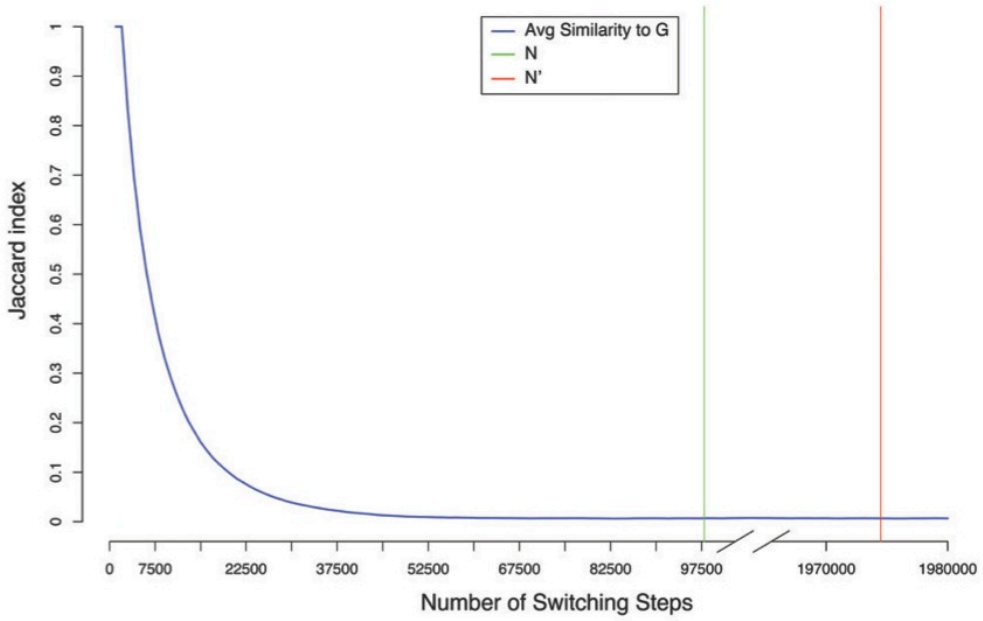
(A): the black curve indicates the difference between the number of edges shared by the original network and its rewired versions at the *k-th* and the *(k-1)-th* switching steps, averaged across 2,500 independent runs of the switching-algorithm. This difference reaches a plateau, close to zero, before *N* switching steps (indicated by the dashed line) and far more before *N'* switching steps (the right limit of the x-axis). The same happens to the Total variation distance and the Kolmogorov distance (respectively blue and red curves) between the distribution of the shared edges at the *k-th* and the *N'-th* switching-step, across the 2,500 independent runs of the switching-algorithm.

(B): evolution of the distribution of common edges between the original network and its rewired version at *k* switching-steps. The color-coding reflects the number of performed switching step. The limit distribution reached at *N'* switching-steps, which is equal to that reached at *N* switching-steps is depicted in black.



Supplementary Figure S4 – Moments of the distribution of common edges between the original network and its rewired versions at *k* switching-steps

The trend of the first 5 moments of the number of edges in common between the original network and its rewired versions in function of the number of switching steps, confirming a convergence time lower than N and far more lower than N'.



Supplementary Figure S5 – Trend of Jaccard similarity across the switching steps for the Breast-cancer dataset

The blue curve indicates the average Jaccard index between the original network and its rewired versions as function of the number of switching steps. The vertical green (respectively red) line indicates our novel bound (respectively the empirical one).

A novel approximated lower bound to the number of switching steps required to rewire bipartite networks: Formal Proof

Here we show how we derived a lower bound N for the number of switching steps (SSs) required by the Switching Algorithm (SA) (see main text of the paper for details)to generate a rewired version of a bipartite network $\mathcal{G}$, providing it with the maximal level of achievable randomness (on average). In what follows, we compute analytically the mean value of the similarity between $\mathcal{G}$ an its rewired version at the k-th switching step $\mathcal{G}^{(k)}$, i.e. $s^{(k)}$ (see Equation 1 defined below). Being the number of edges $|E|$ preserved between $\mathcal{G}$ and $\mathcal{G}^{(k)}$, $s^{(k)}$ is a function of $|E|$ and the number of edges in common between $\mathcal{G}$ and $\mathcal{G}^{(k)}$, i.e. $x^{(k)}$.

We first define a mean-field equation (1) for $x^{(k)}$, which results in a second order recursive function of the number of switching steps. Since this mean-field equation admits a closed form, it is possible to compute from it a unique fixed point $\bar{x}$ and a convergence time N (in terms of number of SSs), once a level of accuracy ε is fixed.

Finally, we prove that the similarity between any pair of rewired versions of $\mathcal{G}$ obtained through different instances of the SA, with N SSs, is lower than their individual similarity to $\mathcal{G}$, hence this algorithm can be used to simulate samples from the uniform distribution of networks with same node set sizes and degree distributions of $\mathcal{G}$. In the following table a scheme of the proof is provided.

1. Computation of the mean-field equation for $x^{(k)}$ and consequently for $s^{(k)}$ (see Lemma 1);

2. Derivation of the fixed point $\bar{x}$ and the convergence time N for the mean-field equation found in Lemma 1 (see Lemma 2);

3. Proof that the SA can be used to simulate samples from the uniform distribution of networks with same node set sizes and degree distributions as $\mathcal{G}$ through N switching steps (see Lemma 3).

Preliminary notation and randomness of a rewired network across switching steps

Let $\mathcal{G} = (\{V_r, V_c\}, E)$ be a bipartite network and $\mathcal{B}$ its $n_r \times n_c$ binary incidence matrix (with $n_r = |V_r|$ and $n_c = |V_c|$), with $|E|$ = number of edges, $V_r = \{1, \dots, n_r\}$ and $V_c = \{1, \dots, n_c\}$. In what follows, we will indicate with **1** (respectively **0**) all the entries of a matrix (or a vector) assuming value 1 (resp. 0). The number of edges in a complete bipartite graph with same node set sizes of $\mathcal{G}$ will be indicated with t $(= n_c n_r)$. The superscript (k) will indicate the observation time (in terms of SSs) of the object under consideration. For example $\mathcal{B}^{(k)}$ will indicate the incidence matrix of the original network $\mathcal{G}$ after k switching steps (i.e. $\mathcal{G}^{(k)}$), $E^{(k)}$ the set of edges of the same network and so on. When referring to the initial network (i.e. $k = 0$) this superscript will be omitted.

The switching algorithm (SA) proceeds through a sequence of switching steps (SS). Let $w_{i,j}$ be the i,j-th entry of $\mathcal{B}^{(k-1)}$ and $\mathcal{L}^{(k-1)}$ the edge-list of $\mathcal{G}^{(k-1)}$, i.e. a $|E| \times 2$ matrix such that $(l_{i,1}^{(k-1)}, l_{i,2}^{(k-1)}) \in E^{(k-1)}, \forall i = 1, \dots, |E|$.

At the k-th SS, the following actions are performed:

1. two numbers m, n are randomly selected, such that $n \neq m$ and $n, m \in \{1, \dots, |E|\}$,
2. the terminal nodes of the corresponding edges are considered,
   $a = l_{n,1}^{(k-1)}$, $b = l_{n,2}^{(k-1)}$, $c = l_{m,1}^{(k-1)}$, $d = l_{m,2}^{(k-1)}$,
3. if $w_{a,d} = 0, w_{c,b} = 0, b \neq c, a \neq d, a \neq b, c \neq d$:
   a. $w_{a,d} = 1, w_{c,b} = 1$,
   b. $w_{a,b} = 0, w_{c,d} = 0$.
   c. $l_{n,2} = d, l_{m,2} = b$.

Let $\mathcal{B}^{(k)}$ be the incidence matrix of $\mathcal{G}^{(k)}$, after k of these steps and $s^{(k)}$ the Jaccard Index (JI) (2) between $\mathcal{B}$ and $\mathcal{B}^{(k)}$. Each switching step does not alter the node degrees of $\mathcal{G}$, the total number of **1**s in $\mathcal{B}$, as well as its row- and column-wise sums. As a consequence, $s^{(k)}$ can be written as:

$$s^{(k)} = \frac{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} w_{i,j}^{(k)} w_{i,j}}{2|E| - \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} w_{i,j}^{(k)} w_{i,j}} = \frac{x^{(k)}}{2|E| - x^{(k)}},$$

Equation 1

where $x^{(k)} \in [0, |E|]$ is equal to the total number of **1**s in the Hadamard product $\mathcal{B} \circ \mathcal{B}^{(k)}$ (i.e. the number ones in common positions across the two adjacency matrices, hence common edges in the two corresponding networks).

Lemma 1

The mean-field equation for $x^{(k+1)}$ is equal to

$$x^{(k+1)} = mx^{(k)} + q = \frac{t|E| - 2t + 2|E|}{|E|t} x^{(k)} + \frac{2t|E| - 2|E|^2}{t^2}.$$

Equation 2

Proof. After a switching step, turning $\mathcal{B}^{(k)}$ into $\mathcal{B}^{(k+1)}$, 5 possible values can be assumed by $x^{(k+1)}$:

1) $x^{(k+1)} = f_1(x^{(k)}) = x^{(k)} + 1$: unitary increment. The switching step is successfully performed and one of the following conditions is verified:

- $(a, b), (c, d) \notin E$ and only one between (a, d) and (c, b) is in E;
- only one between (a, b) and (c, d) is in E and $(a, d), (c, b) \in E$.

2) $x^{(k+1)} = f_2(x^{(k)}) = x^{(k)} - 1$: unitary decrement. The rewiring step is successfully performed and one of the following conditions is verified:

- $(a, b), (c, d) \in E$ and only one between (a, d) and (c, b) is in E;
- only one between (a, b) and (c, d) is in E and $(a, d), (c, b) \notin E$.

3) $x^{(k+1)} = f_3(x^{(k)}) = x^{(k)} + 2$: maximal increment. The rewiring step is successfully performed and $(a, b), (c, d) \notin E$ while $(a, d), (c, b) \in E$.

4) $x^{(k+1)} = f_4(x^{(k)}) = x^{(k)} - 2$: maximal decrement. The rewiring step is successfully performed and $(a, b), (c, d) \in E$ while $(a, d), (c, b) \notin E$.

5) $x^{(k+1)} = f_5(x^{(k)}) = x^{(k)}$: null variation. Otherwise.

Note that E refers to to the edge set of $\mathcal{G}$ (i.e. $E^{(0)}$).

Table 1 contains a summary of the five possible values that $x^{(k+1)}$ can assume.

f_1	f_2	f_3	f_4	f_5
$+1$	-1	$+2$	-2	$+0$

Table 1: Possible variations of $x^{(k+1)}$.

If we indicate with $p_i^{(k)} = P(x^{(k+1)} = f_i(x^{(k)}))$ (i.e. probability of each of the 5 cases, for $i = 1, \dots, 5$), then $x^{(k+1)}$ is equal, on average, to:

$$x^{(k+1)} = \sum_{i=1}^{5} p_i^{(k)} f_i(x^{(k)}).$$

Equation 3

Explicating the probabilities $p_i^{(k)}$ for $i = 1, \dots, 5$ (see Proposition 1, 2 and 3 below and Figure 1) reduces Equation 3 to Equation 2.

□

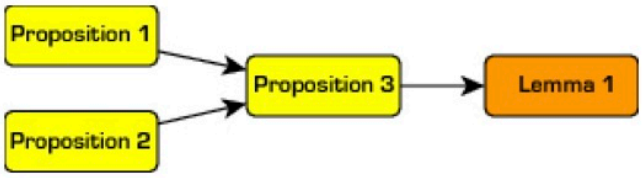


Figure 1 Proof scheme for Lemma 1

In order to prove Proposition 1, 2 and 3, we will make use of the following additional notation.

Consider a, b, c, d defined in 2nd step of the SA and $\mathcal{B}_{i,j}^{(k)}$ the i,j-th element of the incidence matrix of the graph $\mathcal{B}^{(k)}$. With $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^{(k)}$ we will indicate the submatrix of $\mathcal{B}^{(k)}$ corresponding to the four positions $\alpha = b_{a,b}^{(k)}, \delta = b_{c,d}^{(k)}, \beta = b_{a,d}^{(k)}, \gamma = b_{c,b}^{(k)}$. In what follows, when an entry of the $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^{(k)}$ can be neglected then it will be indicated with the $\cdot$ symbol. When $k = 0$ the corresponding submatrix $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^{(0)}$ will be denoted with $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ and it will contain entries from the original matrix $\mathcal{B}$. We will denote the probabilities of the following eight possible events as follows:

$$q_s^{(k)} = P(QS_k^+) = P\left(\begin{pmatrix} 1 & \cdot \\ \cdot & \cdot \end{pmatrix} \middle| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{(k)} \right) = P(QS_k^-) = P\left(\begin{pmatrix} \cdot & \cdot \\ \cdot & 1 \end{pmatrix} \middle| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{(k)} \right)$$

$$p_s^{(k)} = P(PS_k^+) = P\left(\begin{pmatrix} 0 & \cdot \\ \cdot & \cdot \end{pmatrix} \middle| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{(k)} \right) = P(PS_k^-) = P\left(\begin{pmatrix} \cdot & \cdot \\ \cdot & 0 \end{pmatrix} \middle| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{(k)} \right)$$

$$q_f^{(k)} = P(QF_k^+) = P\left(\begin{pmatrix} \cdot & 1 \\ \cdot & \cdot \end{pmatrix} \middle| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{(k)} \right) = P(QF_k^-) = P\left(\begin{pmatrix} \cdot & \cdot \\ 1 & \cdot \end{pmatrix} \middle| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{(k)} \right)$$

$$p_f^{(k)} = P(PF_k^+) = P\left(\begin{pmatrix} \cdot & 0 \\ \cdot & \cdot \end{pmatrix} \middle| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{(k)}\right) = P(PF_k^-) = P\left(\begin{pmatrix} \cdot & \cdot \\ 0 & \cdot \end{pmatrix} \middle| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{(k)}\right)$$

For example, the value $q_s^{(k)}$ is the probability of having $w_{a,b} = 1$ in the initial graph knowing that the rewiring step is performed successfully. The other events and probabilities have similar interpretations.

Proposition 1

The probability p_r of the event $PR =$ 'a rewiring step is successfully performed' is equal to:

$$p_r = P\left(\begin{pmatrix} \cdot & 0 \\ 0 & \cdot \end{pmatrix}^{(k)} \middle| \begin{pmatrix} 1 & \cdot \\ \cdot & 1 \end{pmatrix}^{(k)}\right) \simeq \left(\frac{t - |E|}{t}\right)^2.$$

Equation 4

Proof.

$$p_r = P\left(w_{a,d}^{(k)} = 0 \wedge w_{c,b}^{(k)} = 0\right) = P\left(w_{a,d}^{(k)} = 0\right)P\left(w_{c,b}^{(k)} = 0 \middle| w_{a,d}^{(k)} = 0\right) = \left(\frac{t-|E|}{t}\right)\left(\frac{t-|E|-1}{t-1}\right) \simeq \left(\frac{t-|E|}{t}\right)^2.$$

The probability $P\left(w_{a,d}^{(k)} = 0\right)$ is computed as the ratio between the number of positive cases (available positions) and the number of possible cases (all the positions). Note that this probability does not depend on k.

Proposition 2

In the above notation:

$$q_s^{(k)} \simeq \frac{x^{(k)}}{|E|}, \quad p_s^{(k)} \simeq \frac{|E|-x^{(k)}}{|E|}, \quad q_f^{(k)} \simeq \frac{|E|-x^{(k)}}{t-|E|}, \quad p_f^{(k)} \simeq \frac{t-2|E|+x^{(k)}}{t-|E|}.$$

Proof. Pretend that at the step k there are $x^{(k)}$ ones in common between $\mathcal{B}^{(k)}$ and $\mathcal{B}$, and that $w_{a,b}^{(k)} = 1$, then the probability that in the initial graph $w_{a,b} = 1$ is equal to $\frac{x^{(k)}}{|E|}$ (positive cases divided by possible cases). Similarly, for $q_f^{(k)}$ the possible cases are $t - |E|$, i.e. the number of available positions in which the new non null entry can be placed, and the positive cases are $|E| - x^{(k)}$; then

$$q_f^{(k)} \simeq \frac{|E| - x^{(k)}}{t - |E|},$$

Equation 5

where we made use of following approximations: $x^{(k)} - 1 \simeq x^{(k)}$ and $|E| - 1 \simeq |E|$. The rest of the proof can be deduced observing that $p_s^{(k)} = 1 - q_s^{(k)}$ and $p_f^{(k)} = 1 - q_f^{(k)}$.

Proposition 3

The probabilities $p_i^{(k)}$, $i = 1, \dots, 5$ are equal to:

$$p_1^{(k)} \simeq \frac{2(|E|-x^{(k)})^3(2x^{(k)}+t-2|E|)}{t^2|E|^2}, \qquad p_2^{(k)} \simeq \frac{2(|E|-x^{(k)})(x^{(k)}+t-2|E|)(2x^{(k)}+t-2|E|)x^{(k)}}{t^2|E|^2},$$

$$p_3^{(k)} \simeq \frac{(x^{(k)}-|E|)^4}{t^2|E|^2}, \qquad p_4^{(k)} \simeq \frac{x^{(k)}(x^{(k)}+t-2|E|)^2x^{(k)}}{t^2|E|^2}, \qquad p_5^{(k)} = 1 - p_4^{(k)} - p_3^{(k)} - p_2^{(k)} - p_1^{(k)}.$$

Proof. Using the definition of $f_1(x^{(k)})$ in Lemma 1, the value p_r computed in Proposition 1 and the four probabilities in Proposition 2, it follows that:

$$p_1^{(k)} = P(PR \wedge (((PS_k^+ \wedge PS_k^-) \wedge ((QF_k^+ \wedge PF_k^-) \vee (QF_k^- \wedge PF_k^+))) \vee (((QS_k^+ \wedge PS_k^-) \vee (QS_k^- \wedge PS_k^+)) \wedge (QF_k^+ \wedge QF_k^-)))).$$

This can be rewritten (omitting the probabilities of the prior events, for the sake of simplicity) as:

$$p_1^{(k)} = P\left[\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{(k)} \wedge \left\{\left[\begin{pmatrix}0 & \cdot\\ \cdot & 0\end{pmatrix} \wedge \left[\begin{pmatrix}\cdot & 1\\ 0 & \cdot\end{pmatrix} \vee \begin{pmatrix}\cdot & 0\\ 1 & \cdot\end{pmatrix}\right]\right] \vee \left[\left[\begin{pmatrix}1 & \cdot\\ \cdot & 0\end{pmatrix} \vee \begin{pmatrix}0 & \cdot\\ \cdot & 1\end{pmatrix}\right] \wedge \begin{pmatrix}\cdot & 1\\ 1 & \cdot\end{pmatrix}\right]\right\}\right]$$

$$= P\left[\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{(k)} \wedge \left[\begin{pmatrix}0 & \cdot\\ \cdot & 0\end{pmatrix} \wedge \left[\begin{pmatrix}\cdot & 1\\ 0 & \cdot\end{pmatrix} \vee \begin{pmatrix}\cdot & 1\\ 0 & \cdot\end{pmatrix}\right]\right]\right] + P\left[\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{(k)} \wedge \left[\left[\begin{pmatrix}1 & \cdot\\ \cdot & 0\end{pmatrix} \vee \begin{pmatrix}0 & \cdot\\ \cdot & 1\end{pmatrix}\right] \wedge \begin{pmatrix}\cdot & 1\\ 1 & \cdot\end{pmatrix}\right]\right]$$

$$\simeq p_r\left[p_s^2(1 - p_f^2 - q_f^2) + (1 - p_s^2 - q_s^2)q_f^2\right].$$

Similarly:

$$p_2^{(k)} = P\left[\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{(k)} \wedge \left\{\left[\begin{pmatrix}1 & \cdot\\ \cdot & 1\end{pmatrix} \wedge \left[\begin{pmatrix}\cdot & 1\\ 0 & \cdot\end{pmatrix} \vee \begin{pmatrix}\cdot & 0\\ 1 & \cdot\end{pmatrix}\right]\right] \vee \left[\left[\begin{pmatrix}1 & \cdot\\ \cdot & 0\end{pmatrix} \vee \begin{pmatrix}0 & \cdot\\ \cdot & 1\end{pmatrix}\right] \wedge \begin{pmatrix}\cdot & 0\\ 0 & \cdot\end{pmatrix}\right]\right\}\right]$$

$$= P\left[\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{(k)} \wedge \left[\begin{pmatrix}1 & \cdot\\ \cdot & 1\end{pmatrix} \wedge \left[\begin{pmatrix}\cdot & 1\\ 0 & \cdot\end{pmatrix} \vee \begin{pmatrix}\cdot & 0\\ 1 & \cdot\end{pmatrix}\right]\right]\right] + P\left[\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{(k)} \wedge \left[\left[\begin{pmatrix}1 & \cdot\\ \cdot & 0\end{pmatrix} \vee \begin{pmatrix}0 & \cdot\\ \cdot & 1\end{pmatrix}\right] \wedge \begin{pmatrix}\cdot & 0\\ 0 & \cdot\end{pmatrix}\right]\right]$$

$$\simeq p_r\left[q_s^2(1 - p_f^2 - q_f^2) + (1 - p_s^2 - q_s^2)p_f^2\right],$$

$$p_3^{(k)} = P\left[\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{(k)} \wedge \begin{pmatrix}0 & \cdot\\ \cdot & 0\end{pmatrix} \wedge \begin{pmatrix}\cdot & 1\\ 1 & \cdot\end{pmatrix}\right] \simeq p_r p_s^2 q_f^2,$$

$$p_4^{(k)} = P\left[\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{(k)} \wedge \begin{pmatrix}1 & \cdot\\ \cdot & 1\end{pmatrix} \wedge \begin{pmatrix}\cdot & 0\\ 0 & \cdot\end{pmatrix}\right] \simeq p_r q_s^2 p_f^2.$$

Combining Equation 1 and Equation 2, the mean-field equation for the JI $s^{(k+1)}$ can be written as:

$$s^{(k+1)} = \frac{2x^{(k)}t^2 - x^{(k)}t^2|E| - 2x^{(k)}t|E| - 2t|E|^2 + 2|E|^3}{x^{(k)}t^2|E| - 2x^{(k)}t^2 + 2x^{(k)}t|E| - 2t^2|E|^2 + 2t|E|^2 - 2|E|^3}.$$

Equation 6

The mean-field Equation 6 is an approximation because Equation 5 does not consider the preservation of the degree distributions. To take this constraint into account, we slightly modify Equation 5 as follows:

$$q_f^{(k)} \simeq \frac{|E| - x^{(k)}}{t - |E| - z},$$

Equation 7

where $t - |E| - z$ represents the number of *available positions* where the new non null entry can be placed. The value z depends on the initial graph $\mathcal{G}$ and is related to the admissible configurations of the BEM keeping constant the degree distributions. In Lemma 2 we show that this value can be neglected, or better, that the number of SSs required to approach the fixed point is maximum for $z = 0$.

If reformulating Proposition 1, 2, 3 and Lemma 1 according to this modification the mean-field equation for $x^{(k+1)}$ is equal to:

$$x^{(k+1)} = m(z)x^{(k)} + q(z) =$$

$$= \frac{(|E|-2)t^3 - [(|E|-2)z - 4|E| + |E|^2]t^2 - 2(2z|E| + |E|^2)t + 2z|E|^2}{t^3|E| - (z|E| + |E|^2)t^2} x^{(k)}$$

$$+ \frac{2t^2|E|^2 - 4t|E|^3 + 2|E|^4}{t^3|E| - (z|E| + |E|^2)t^2}.$$

Equation 8

The demonstration of Lemma 2 follows from Lemma 1, Proposition 4 and Proposition 5 (as summarized in Figure 2).

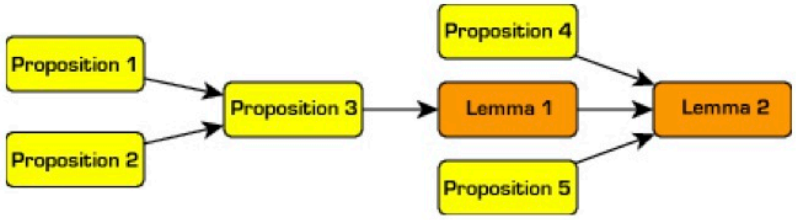


Figure 2: Proof scheme for Lemma 2

Proposition 4

The unique fixed point $\overline{x}$ of Equation 8 is:

$$\overline{x} = \frac{|E|^2}{t - z}.$$

Equation 9

Proof. Let us solve $x^{(k+1)} = x^{(k)} = \overline{x}$:

$$0 = p_1(\overline{x})(\overline{x}+1) + p_2(\overline{x})(\overline{x}-1) + p_3(\overline{x})(\overline{x}+2) + p_4(\overline{x})(\overline{x}-2) + p_5(\overline{x})\overline{x} - \overline{x}$$

$$= p_1(\overline{x}) - p_2(\overline{x}) + 2p_3(\overline{x}) - 2p_4(\overline{x})$$

$$= \frac{-2(t-|E|)^2(\overline{x}z - \overline{x}t + |E|^2)}{(z - t + |E|)|E|t^2}$$

$$= \overline{x}z - \overline{x}t + |E|^2$$

$$\rightarrow \overline{x} = \frac{|E|^2}{t - z}$$

Proposition 5

For a fixed real $\varepsilon \leq 1$, $|x^{(k)} - \overline{x}| < \varepsilon$ for all $k > N$ with

$$N = \log_{m(z)} g(z, \varepsilon) \qquad with \qquad g(z, \varepsilon) = \frac{\varepsilon(t-z)}{(t-|E|-z)|E|}.$$

Proof. From Equation 8 it follows that:

$$\begin{aligned} x^{(k+1)} &= m(z)x^{(k)} + q(z) \\ &= (m(z)+1)x^{(k)} - m(z)x^{(k-1)}, \end{aligned}$$

which is a second − order linear recursive sequence admitting

$$F(x) = x^2 - (m(z)+1)x + m(z)$$

as characteristic polynomial. As shown in (3) we can write

$$x^{(k+1)} = ar^{k+1} + bs^{k+1},$$

where r and s are the two roots of F and a and b are constants

$$= am(z)^{k+1} + b, \qquad in\ our\ case\ r = m(z), \qquad s = 1,$$

$$= \left(|E| - \frac{q(z)}{1-m(z)}\right) m(z)^{k+1} + \frac{q(z)}{1-m(z)}$$

Equation 10

$given\ that\ x^{(0)} = |E|\ and\ x^{(1)} = m(z)|E| + q(z).$

Fixed $\varepsilon \leq 1$,

$$|x^{(N)} - \overline{x}| < \varepsilon \Leftrightarrow \left|\left(|E| - \frac{q(z)}{1-m(z)}\right) m(z)^k\right| < \varepsilon \Leftrightarrow$$

$$N > \log_{m(z)} g(z, \varepsilon) \qquad with \qquad g(z, \varepsilon) = \frac{\varepsilon(t-z)}{(t-|E|-z)|E|}.$$

Equation 11

Since $0 < m(z) \leq 1$ the previous inequality holds.

□

Lemma 2

Let d denote the edge density of $\mathcal{G}$, namely $d = \frac{|E|}{t} \in [0,1]$ and $\varepsilon = 1$, then

$$N \simeq \frac{|E|}{2(1-d)} \ln(|E| - d|E|).$$

Equation 12

Proof. Since $m'(z) = -\frac{2(t-|E|)^2}{(t(t-s-|E|))^2} < 0$ and $\frac{\partial}{\partial z} g(z,\varepsilon) = -\frac{(t-z)^2}{e^2} < 0$, the maximum value for N of Equation 11 is reached for $z = 0$ and its value is:

$$\begin{aligned} N &= \log_{\frac{(|E|-2)t+2|E|}{|E|t}} \frac{t}{t|E|-|E|^2} \\ &= \log_{1-\frac{2(1-d)}{dt}} \frac{1}{|E|-d|E|} \\ &= \frac{\ln \frac{1}{|E|-d|E|}}{\ln 1-\frac{2(1-d)}{dt}} \\ &\sim \frac{dt}{2(1-d)} \ln(|E| - d|E|) \quad using\ \ln[1+x] \sim x\ for\ |x| < 1 \\ &= \frac{|E|}{2(1-d)} \ln(|E| - d|E|). \end{aligned}$$

□

Pairwise-similarity

Let $r^{(k)} = s(\mathcal{B}^{(k)}, \mathcal{C}^{(k)})$ where $\mathcal{B}^{(k)}$ and $\mathcal{C}^{(k)}$ are the incidence matrices of two rewired version of $\mathcal{G}$, obtained through the SA with k SSs . In this section we will show that the similarity between any pair of rewired versions of $\mathcal{G}$ obtained through different instances of the SA, with k SSs, is lower than their individual similarity to $\mathcal{G}$.

Proposition 6

Using the same notation and Proposition 1 and 2, with $z = 0$ it follows that:

$$r^{(k+1)} = \overline{m}r^{(k)} + \overline{q} = \frac{(|E| - 4)t^3 - (|E|^2 - 8e)t^2 - 4|E|^2 t}{t^3|E| - |E|^2 t^2} r^{(k)} - 4\frac{(t^2|E| - 2t|E|^2 + |E|^3)}{|E|t^2 - t^3}.$$

Equation 13

Proof. Similarly to the proof of Lemma 1 the value $r^{(k+1)}$ can be estimated as:

$$r^{(k+1)} = \sum_{i=1}^{9} q_i^{(k)} g_i\left(r^{(k)}\right),$$

Equation 14

where the values of g_i are summarized in Table 2. For more details see Appendix A.

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
$+4$	-4	$+3$	-3	$+2$	-2	$+1$	-1	$+0$

Table 2: Possible variations for $r^{(k+1)}$.

The rest of proof follows from the explication of the probabilities $q_i^{(k)}$ $i = 1, \ldots, 9$ (Proposition 6 and Proposition 7 as summarized in Figure 3).

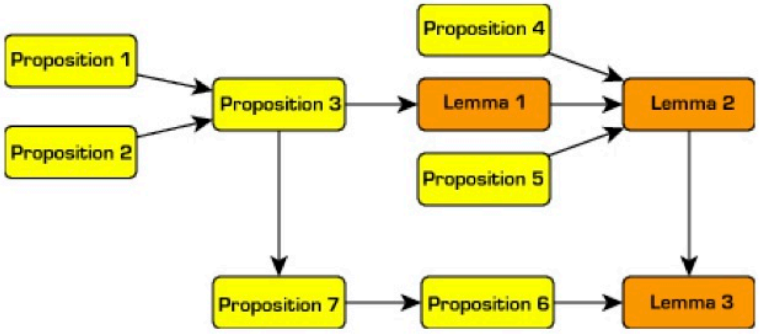


Figure 3: Proof scheme for Lemma 3

Proposition 7

From the definition of the probabilities in Proposition 3 we can compute $q_1^{(k)}, i = 1, \dots, 9$ (values and details in Appendix B).

Proof. Similarly to the proof of Lemma 1:

$$q_1^{(k)} = P\big(PR_b \wedge PR_c \wedge (PS_b^+ \wedge PS_b^- \wedge PS_b^+ \wedge PS_b^-) \wedge (QF_c^+ \wedge QF_c^- \wedge QF_c^+ \wedge QF_c^-)\big) \sim p_r^2 p_s^4 q_f^4,$$

$$q_2^{(k)} = P\big(PR_b \wedge PR_c \wedge (QS_b^+ \wedge QS_b^- \wedge QS_b^+ \wedge QS_b^-) \wedge (PF_c^+ \wedge PF_c^- \wedge PF_c^+ \wedge PF_c^-)\big) \sim p_r^2 q_s^4 p_f^4,$$

$$\begin{aligned} q_3^{(k)} = P(&PR_b \wedge PR_c \wedge (\\ &((PS_b^+ \wedge PS_b^- \wedge PS_b^+ \wedge PS_b^-) \wedge \\ &((QF_c^+ \wedge QF_c^- \wedge QF_c^+ \wedge PF_c^-) \vee (QF_c^+ \wedge QF_c^- \wedge PF_c^+ \wedge QF_c^-) \vee \\ &(QF_c^+ \wedge PF_c^- \wedge QF_c^+ \wedge QF_c^-) \vee (PF_c^+ \wedge QF_c^- \wedge QF_c^+ \wedge QF_c^-)) \\ &) \vee (\\ &((PS_c^+ \wedge PS_c^- \wedge PS_c^+ \wedge QS_c^-) \vee (PS_c^+ \wedge PS_c^- \wedge QS_c^+ \wedge PS_c^-) \vee \\ &(PS_c^+ \wedge QS_c^- \wedge Ps_c^+ \wedge PS_c^-) \vee (QS_c^+ \wedge PS_c^- \wedge PS_c^+ \wedge PS_c^-)) \wedge \\ &(QF_b^+ \wedge QF_b^- \wedge QF_b^+ \wedge QF_b^-)))) \sim p_r^2 (4p_s^4 q_f^3 p_f + 4p_s^3 q_s q_f^4). \end{aligned}$$

Similarly:

$$q_4^{(k)} \sim p_r^2 (4q_s^4 q_f p_f^3 + 4q_s^3 p_s p_f^4).$$

$$q_5^{(k)} \sim p_r^2(6q_f^2p_f^2p_s^4 + 16q_f^3p_fp_s^3q_s + 6q_f^4p_s^2q_s^2) + 2pr(1-pr)(q_f^2p_s^2).$$

$$q_6^{(k)} \sim p_r^2(6q_f^2p_f^2q_s^4 + 16p_f^3q_fq_s^3p_s + 6p_f^4p_s^2q_s^2) + 2p_r(1-p_r)(p_f^2q_s^2).$$

$$q_7^{(k)} \sim p_r^2(4q_fp_f^3p_s^4 + 24p_f^2q_f^2p_s^3q_s + 24p_fq_f^3p_s^2q_s^2 + 4q_f^4p_sq_s^3) + 2p_r(1-p_r)(2p_s^2q_fp_f + 2p_sq_sq_f^2).$$

$$q_8^{(k)} \sim p_r^2(4q_f^3p_fq_s^4 + 24q_f^2p_f^2q_s^3p_s + 24q_fp_f^3p_s^2q_s^2 + 4p_f^4q_sp_s^3) + 2pr(1-pr)(2q_s^2p_fq_f + 2p_sq_sp_f^2).$$

$$q_9^{(k)} = 1 - \sum_{i=1}^{8} q_i^{(k)}.$$

Lemma 3

Let be $x^{(k)}$ defined as in Proposition 5 and $z = 0$, assuming $|E| > 6$, then the fixed point $\bar{r}$ of Equation 12 is

$$\bar{r} = \frac{|E|^2}{t}$$

and for all $k = 1, \dots, N$, follows that:

$$r^{(k)} \leq x^{(k)}.$$

Proof. From Equation 12, $\bar{r}$ is a fixed point if and only if:

$$0 = r^{(k+1)} - r^{(k)} = \frac{-4(t-|E|)^2(|E|^2 - r^{(k)}t)}{|E|(|E|-t)t^2}.$$

The unique admissible root of this equivalence is $\frac{|E|^2}{t}$.

The sequence in Equation 12 is again a second order linear sequence for which a closed form can be computed as shown in (3):

$$r^{(k)} = \frac{t|E| - |E|^2}{t}\left(\frac{(|E|-4)t + 4|E|}{t^2}\right)^k + \frac{|E|^2}{t} \qquad and$$

$$x^{(k)} = \frac{t|E| - |E|^2}{t}\left(\frac{(|E|-2)t + 2|E|}{|E|t}\right)^k + \frac{|E|^2}{t} \qquad so$$

$$r^{(k)} \leq x^{(k)} \Leftrightarrow \frac{(|E|-4)t + 4|E|}{t^2} \leq \frac{(|E|-2)t + 2|E|}{|E|t}$$

$$\Leftrightarrow -\frac{(|E|-2)t^2 + 4t|E| - 4|E|^2}{|E|t^2} \leq 0$$

$$\Leftrightarrow -(|E|-2)t^2 - 4t|E| + 4|E|^2 \leq 0$$

$$\Leftrightarrow 4|E|^2 \leq (|E|-2)t^2 + 4t|E|.$$

Equation 15

Assuming that $|E| > 6$, the last inequality is always satisfied because:

$$4t^2 \leq (|E|-2)t^2 + 4t|E| \Leftarrow (6-|E|)t \leq 4|E|.$$

In conclusion $r^{(k)} \leq x^{(k)}$.

Bound generalization

Binary event matrices coding for large-scale cancer genomic datasets tend to be sparse (i.e. the number of variants is small compared with the product between the number of sequenced genes and the number of samples). Additionally, usually few genes (i.e. oncogenes and tumor-suppressor genes) are altered in a large number of samples whereas a large amount of genes is altered in few samples.

As a consequence, when coded as bi-partite network such a dataset results into a low edge-density, scale-free network. This allows the probability of the event $PR =$ "a switching step is succesfully performed", p_r to be approximated as we have done in Proposition 1.

In the case of datasets yielding networks with a very high edge density and/or a high level of homophily, then this probability cannot be computed as in the previous case and a more general bound for the number of steps would be

$$N = \frac{|E|(1-d)}{2p_r}\ln|E|(1-d).$$

In this case a lower bound for the minimum number of **successful** switching steps to be performed would be

$$M = \frac{|E|(1-d)}{2}\ln|E|(1-d)$$

Obviously, in the cases where the effective value p_r can be computed or estimated, the number of switching steps directly follows from the equation $N = \frac{M}{p_r}$.

References

1. Barabási AL, Albert R, Jeong H. Mean-field theory for scale-free random networks. Physica A 1999; 272:173-187
2. Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. 1901.
3. Brousseau A. Linear Recursion and Fibonacci Sequences. 1971.

Fast randomisation of large genomic datasets while preserving alteration counts

Andrea Gobbi[1,5,*], Francesco Iorio[2,3,*,†], Kevin J. Dawson[3], David C. Wedge[3], David Tamborero[4], Ludmil B. Alexandrov[3], Nuria Lopez-Bigas[4], Mathew J. Garnett[3], Giuseppe Jurman[1], Julio Saez-Rodriguez[2]

[1]Fondazione Bruno Kessler, Trento – Italy, [2]European Molecular Biology Laboratory – European Bioinformatics Institute, Cambridge – UK, [3]Wellcome Trust Sanger Institute, Cambridge – UK, [4]Universitat Pompeu Fabra, Barcelona – Spain, [5]University of Trento, Trento – Italy.

Appendix A

Using the letters a, b, c, d for $\mathcal{B}^{(k)}$ and $\alpha, \beta, \gamma, \delta$ for $\mathcal{C}^{(k)}$ and introducing $F^{(k)}$, the set of the common edges between $\mathcal{B}^{(k)}$ and $\mathcal{C}^{(k)}$ we have:

1) $g_1(r^{(k)}) = r^{(k)} + 4$: we gain four ones. The two rewiring steps are performed (one for $\mathcal{B}^{(k)}$ and one for $\mathcal{C}^{(k)}$) and $(a, b), (c, d), (\alpha, \beta), (\gamma, \delta) \notin F^{(k)}$ and $(a, d), (c, b), (\alpha, \delta), (\gamma, \beta) \in F^{(k)}$

2) $g_2(r^{(k)}) = r^{(k)} - 4$: we lose four ones. The two rewiring steps are performed and $(a, b), (c, d), (\alpha, \beta), (\gamma, \delta) \in F^{(k)}$ and $(a, d), (c, b), (\alpha, \delta), (\gamma, \beta) \notin F^{(k)}$

3) $g_3(r^{(k)}) = r^{(k)} + 3$: we gain three ones. The two rewiring steps are performed and:
   - $(a, b), (c, d), (\alpha, \beta), (\gamma, \delta) \notin F^{(k)}$ and only three among $(a, d), (c, b), (\alpha, \delta), (\gamma, \beta)$ are elements of $F^{(k)}$ or
   - One among $(a, b), (c, d), (\alpha, \beta), (\gamma, \delta)$ is in $F^{(k)}$ and $(a, d), (c, b), (\alpha, \delta), (\gamma, \beta) \in F^{(k)}$

4) $g_4(r^{(k)}) = r^{(k)} - 3$: we lose three ones. The two rewiring steps are performed and:
   - $(a, b), (c, d), (\alpha, \beta), (\gamma, \delta) \in F^{(k)}$ and only one among $(a, d), (c, b), (\alpha, \delta), (\gamma, \beta)$ is a element of $F^{(k)}$ or
   - Three among $(a, b), (c, d), (\alpha, \beta), (\gamma, \delta)$ are in $F^{(k)}$ and $(a, d), (c, b), (\alpha, \delta), (\gamma, \beta) \notin F^{(k)}$

5) $g_5(r^{(k)}) = r^{(k)} + 2$: we gain two ones.

- The two rewiring steps are performed and:
  - ◊ $(a,b),(c,d),(\alpha,\beta),(\gamma,\delta) \notin F^{(k)}$ and only two among $(a,d),(c,b),(\alpha,\delta),(\gamma,\beta)$ are elements of $F^{(k)}$
  - ◊ $(a,b) \in F^{(k)}$ (or one of the other) and
    - * if $(a,d) \in F^{(k)}$ two among $(c,b),(\alpha,\delta),(\gamma,\beta)$ are in $F^{(k)}$
    - * if $(a,d) \notin F^{(k)}$ $(c,b),(\alpha,\delta),(\gamma,\beta) \in F^{(k)}$
  - ◊ $(a,b)(c,d) \in F^{(k)}$ (or any other couple) and $(a,d),(c,b),(\alpha,\delta),(\gamma,\beta) \in F^{(k)}$
- Only one of the two rewiring steps are performed (let say $\mathcal{B}^{(k)}$) and:
  - ◊ $(a,b),(c,d) \notin F^{(k)}$ and $(a,d),(c,b) \in F^{(k)}$

6) $g_6(r^{(k)}) = r^{(k)} - 2$: we lose two ones.

- The two rewiring steps are performed and:
  - ◊ $(a,b),(c,d),(\alpha,\beta),(\gamma,\delta) \in F^{(k)}$ and only two among $(a,d),(c,b),(\alpha,\delta),(\gamma,\beta)$ are elements of $F^{(k)}$
  - ◊ $(a,b) \notin F^{(k)}$ (or one of the other) and
    - * if $(a,d) \notin F^{(k)}$ one among $(c,b),(\alpha,\delta),(\gamma,\beta)$ is in $F^{(k)}$
    - * if $(a,d) \in F^{(k)}$ $(c,b),(\alpha,\delta),(\gamma,\beta) \notin F^{(k)}$
  - ◊ $(a,b)(c,d) \notin F^{(k)}$ (or any other couple) and $(a,d),(c,b),(\alpha,\delta),(\gamma,\beta) \notin F^{(k)}$
- Only one of the two rewiring steps are performed (let say $\mathcal{B}^{(k)}$) and:
  - ◊ $(a,b),(c,d) \in F^{(k)}$ and $(a,d),(c,b) \notin F^{(k)}$

7) $g_7(r^{(k)}) = r^{(k)} + 1$: we gain a one.

- The two rewiring steps are performed and:
  - ◊ $(a,b),(c,d),(\alpha,\beta),(\gamma,\delta) \notin F^{(k)}$ and only one among $(a,d),(c,b),(\alpha,\delta),(\gamma,\beta)$ is an element of $F^{(k)}$
  - ◊ $(a,b) \in F^{(k)}$ (or one of the other) and
    - * if $(a,d) \in F^{(k)}$ one among $(c,b),(\alpha,\delta),(\gamma,\beta)$ is in $F^{(k)}$
    - * if $(a,d) \notin F^{(k)}$ two among $(c,b),(\alpha,\delta),(\gamma,\beta)$ are in $F^{(k)}$
  - ◊ $(a,b)(c,d) \in F^{(k)}$ (or any other couple) and
    - * if $(a,d),(c,b) \in F^{(k)}$ one among $(\alpha,\delta),(\gamma,\beta)$ is in $F^{(k)}$
    - * if $(a,d) \in F^{(k)}$ and $(c,b) \notin F^{(k)}$ (or viceversa) $(\alpha,\delta),(\gamma,\beta) \in F^{(k)}$
  - ◊ Three among $(a,b),(c,d),(\alpha,\beta),(\gamma,\delta)$ are in $F^{(k)}$ and $(a,d),(c,b),(\alpha,\delta),(\gamma,\beta) \in F^{(k)}$
- Only one of the two rewiring steps are performed (let say $\mathcal{B}^{(k)}$) and:
  - ◊ $(a,b),(c,d) \notin F^{(k)}$ and only one among (a,d) and (c,b) is an element of $F^{(k)}$ or
  - ◊ $(a,b) \in F^{(k)}$,$(c,d) \notin F^{(k)}$ and $(a,d),(c,b) \in F^{(k)}$

8) $g_8(r^{(k)}) = r^{(k)} - 1$: we lose a one.

- The two rewiring steps are performed and:
  - ◊ $(a,b),(c,d),(\alpha,\beta),(\gamma,\delta) \in F^{(k)}$ and three among $(a,d),(c,b),(\alpha,\delta),(\gamma,\beta)$ are elements of $F^{(k)}$
  - ◊ One among $(a,b),(c,d),(\alpha,\beta),(\gamma,\delta)$ is in $F^{(k)}$ and $(a,d),(c,b),(\alpha,\delta),(\gamma,\beta) \notin F^{(k)}$
  - ◊ $(a,b)(c,d) \in F^{(k)}$ (or any other couple) and
    - * if $(a,d),(c,b) \notin F^{(k)}$ one among $(\alpha,\delta),(\gamma,\beta)$ is in $F^{(k)}$

        * if $(a,d) \in F^{(k)}$ and $(c,b) \notin F^{(k)}$ (or vice versa) $(\alpha,\delta),(\gamma,\beta) \notin F^{(k)}$

    ◊ $(a,b) \notin F^{(k)}$ (or one of the other) and
        * if $(a,d) \notin F^{(k)}$ two among $(c,b),(\alpha,\delta),(\gamma,\beta)$ are in $F^{(k)}$
        * if $(a,d) \notin F^{(k)}$ one among $(c,b),(\alpha,\delta),(\gamma,\beta)$ is in $F^{(k)}$

- Only one of the two rewiring steps are performed (let say $\mathcal{B}^{(k)}$) and:
    ◊ $(a,b),(c,d) \in F^{(k)}$ and only one among (a,d) and (c,b) is an element of $F^{(k)}$ or
    ◊ $(a,b) \in F^{(k)}$,$(c,d) \notin F^{(k)}$ and $(a,d),(c,b) \notin F^{(k)}$

9) $g_9(r^{(k)}) = r^{(k)}$: no variation.

Appendix B

The nine probabilities in Proposition 7 are defined as:

$$q_1^{(k)} \sim \left(\frac{(t-|E|)(|E|-r^{(k)})^2}{(|E|-t)|E|t}\right)^4$$

$$q_2^{(k)} \sim \left(\frac{(t-|E|)(r^{(k)}+t-2|E|)r^{(k)}}{(|E|-t)|E|t}\right)^4$$

$$q_3^{(k)} \sim \frac{4(t-|E|)^4(|E|-r^{(k)})^7(t+2r^{(k)}-2|E|)}{((|E|-t)|E|t)^4}$$

$$q_4^{(k)} \sim \frac{4(t-|E|)^4(|E|-r^{(k)})\left(r^{(k)}(t+r^{(k)}-2|E|)\right)^3(2r^{(k)}+t-2|E|)}{((|E|-t)|E|t)^4}$$

$$q_5^{(k)} \sim \left(\frac{t-|E|}{t}\right)^4\left[6\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^2\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^2\left(\frac{|E|-r^{(k)}}{|E|}\right)^4+\right.$$

$$16\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^3\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)\left(\frac{|E|-r^{(k)}}{|E|}\right)^3\left(\frac{r^{(k)}}{|E|}\right)$$

$$\left.+6\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^4\left(\frac{|E|-r^{(k)}}{|E|}\right)^2\left(\frac{r^{(k)}}{|E|}\right)^2\right]+$$

$$2\left(\frac{t-|E|}{t}\right)^2\left[1-\left(\frac{t-|E|}{t}\right)^2\right]\left[\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^2\left(\frac{|E|-r^{(k)}}{|E|}\right)^2\right]$$

$$q_6^{(k)} \sim \left(\frac{t-|E|}{t}\right)^4\left[6\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^2\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^2\left(\frac{r^{(k)}}{|E|}\right)^4+\right.$$

$$16\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^3\left(\frac{|E|-r^{(k)}}{t-|E|}\right)\left(\frac{r^{(k)}}{|E|}\right)^3\left(\frac{|E|-r^{(k)}}{|E|}\right)$$

$$\left.+6\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^4\left(\frac{|E|-r^{(k)}}{|E|}\right)^2\left(\frac{r^{(k)}}{|E|}\right)^2\right]+$$

$$2\left(\frac{t-|E|}{t}\right)^2\left[1-\left(\frac{t-|E|}{t}\right)^2\right]\left[\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^2\left(\frac{r^{(k)}}{|E|}\right)^2\right]$$

$$q_7^{(k)} \sim \left(\frac{t-|E|}{t}\right)^4\left[4\left(\frac{|E|-r^{(k)}}{t-|E|}\right)\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^3\left(\frac{|E|-r^{(k)}}{|E|}\right)^4+\right.$$

$$24\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^2\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^2\left(\frac{|E|-r^{(k)}}{|E|}\right)^3\left(\frac{r^{(k)}}{|E|}\right)$$

$$\left.+24\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^3\left(\frac{|E|-r^{(k)}}{|E|}\right)^2\left(\frac{r^{(k)}}{|E|}\right)^2+4\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^4\left(\frac{|E|-r^{(k)}}{|E|}\right)\left(\frac{r^{(k)}}{|E|}\right)^3\right]$$

$$+2\left(\frac{t-|E|}{t}\right)^2\left[1-\left(\frac{t-|E|}{t}\right)^2\right]\left[2\left(\frac{|E|-r^{(k)}}{|E|}\right)^2\left(\frac{|E|-r^{(k)}}{t-|E|}\right)\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)+\right.$$

$$\left.2\left(\frac{|E|-r^{(k)}}{|E|}\right)\left(\frac{r^{(k)}}{|E|}\right)\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^2\right]$$

$$q_8^{(k)} \sim \left(\frac{t-|E|}{t}\right)^4 \left[4\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^3\left(\frac{r^{(k)}}{|E|}\right)^4 + 24\left(\frac{|E|-r^{(k)}}{t-|E|}\right)^2\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^2\left(\frac{r^{(k)}}{|E|}\right)^3\left(\frac{|E|-r^{(k)}}{|E|}\right) + 24\left(\frac{|E|-r^{(k)}}{t-|E|}\right)\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^3\left(\frac{r^{(k)}}{|E|}\right)^2\left(\frac{|E|-r^{(k)}}{|E|}\right)^2 + 4\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^4\left(\frac{r^{(k)}}{|E|}\right)\left(\frac{|E|-r^{(k)}}{|E|}\right)^3\right] + 2\left(\frac{t-|E|}{t}\right)^2\left[1-\left(\frac{t-|E|}{t}\right)^2\right]\left[2\left(\frac{r^{(k)}}{|E|}\right)^2\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)\left(\frac{|E|-r^{(k)}}{t-|E|}\right) + 2\left(\frac{r^{(k)}}{|E|}\right)\left(\frac{|E|-r^{(k)}}{|E|}\right)\left(\frac{t-2|E|+r^{(k)}}{t-|E|}\right)^2\right]$$

$$q_9^{(k)} \sim 1 - \sum_{i=1}^{8} q^{(k)}$$

A key point for the calculation of these probabilities is that the number of admissible configurations should be correctly enumerated. As an example, to compute $q_7^{(k)}$, i.e. probability of unitary increment, factors 4 and 24 are defined considering that if originally all the four selected edges are not in $F^{(k)}$ we gain a one if and only if one of the rewired edges is in the original network, and there are exactly 4 possible configurations. If we are in the second of the subcases of the first case, then the possible configurations are summarized in Figure 1.

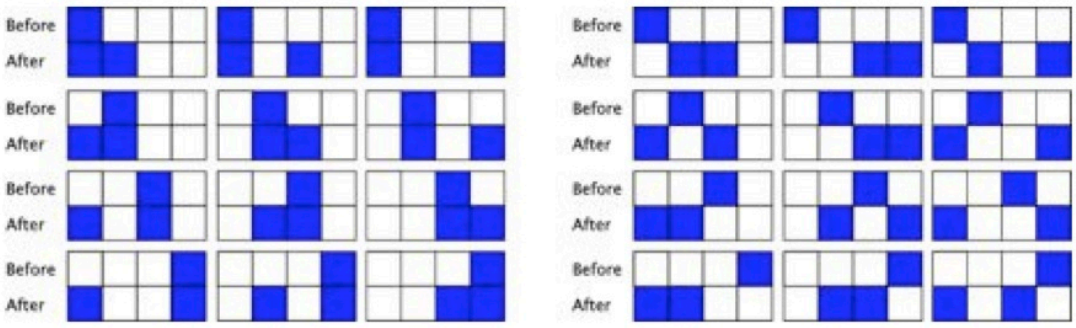


Figure 1: All the 24 configurations resulting in a unitary increment in the second of the subcases of the first case, in the definition of $g_7(r^{(k)})$. The four columns represent the edges and the blue boxes are elements of $F^{(k)}$. The first row of each block represents the configuration before the SS, while the second one represents it after the SS.

GENES	COVERAGE	EXPECTATION [N]	EXPECTATION [Nemp]	SD [N]	SD [Nemp]	p [N]	p [Nemp]	% fdr [N]	% fdr [Nemp]
PIK3CA, TP53	147	138.15	138.08	2.61	2.63	3.53E-04	3.46E-04	8.4	8.2
KRAS, NRAS	117	110.79	110.77	1.97	1.96	7.97E-04	7.36E-04	9.3	8.7
APC, MECOM	174	172.36	172.36	0.55	0.54	1.34E-03	1.16E-03	9.3	8.8
AKAP9, TP53	130	126.48	126.47	1.19	1.19	1.57E-03	1.49E-03	9.3	8.8
ATM, TP53	133	128.55	128.57	1.58	1.6	2.49E-03	2.77E-03	10.1	11
BRAF, KRAS	110	106.14	106.12	1.38	1.39	2.55E-03	2.63E-03	10.1	11
SOX9, TP53	131	127.3	127.32	1.39	1.36	3.82E-03	3.45E-03	13	11.7
FAM123B, TP53	135	130.19	130.21	1.84	1.83	4.48E-03	4.35E-03	13.3	12.9
APC, SMAD2	178	174.51	174.49	1.38	1.39	5.62E-03	5.74E-03	13.3	13.6
ARID1A, TP53	132	128.12	128.14	1.52	1.51	5.39E-03	5.38E-03	13.3	13.6
APC, BRAF	176	173.44	173.44	1.07	1.07	8.11E-03	8.06E-03	17.5	17.4
APC, CDC73	174	172.53	172.54	0.66	0.66	1.27E-02	1.37E-02	22.4	21.6
APC, CNOT1	174	172.54	172.53	0.66	0.66	1.35E-02	1.30E-02	22.4	21.6
APC, DIS3	174	172.54	172.53	0.66	0.66	1.38E-02	1.32E-02	22.4	21.6
APC, MED24	174	172.54	172.52	0.67	0.65	1.42E-02	1.18E-02	22.4	21.6
KRAS, TP53	174	166.53	166.58	3.53	3.47	1.71E-02	1.62E-02	23.6	23.7
CREBBP, TP53	127	125.24	125.24	0.85	0.85	1.89E-02	1.92E-02	23.6	23.7
PPP2R1A, TP53	127	125.25	125.24	0.84	0.85	1.89E-02	1.88E-02	23.6	23.7
TP53, TP53BP1	127	125.23	125.24	0.84	0.86	1.76E-02	2.00E-02	23.6	23.7
CASP8, TP53	127	125.24	125.24	0.86	0.85	2.05E-02	1.92E-02	24.3	23.7
TGFBR2, TP53	130	127.31	127.27	1.36	1.36	2.43E-02	2.27E-02	27.4	25.6
ACSL6, TP53	126	124.82	124.82	0.69	0.7	4.42E-02	4.60E-02	35.6	34.8
ATRX, TP53	126	124.82	124.83	0.7	0.7	4.52E-02	4.69E-02	35.6	34.8
NR2F2, TP53	126	124.83	124.83	0.7	0.7	4.63E-02	4.70E-02	35.6	34.8
NUP107, TP53	126	124.82	124.82	0.7	0.69	4.75E-02	4.35E-02	35.6	34.8
PTEN, TP53	126	124.84	124.82	0.7	0.69	4.81E-02	4.52E-02	35.6	34.8
TCF12, TP53	126	124.83	124.82	0.69	0.7	4.54E-02	4.46E-02	35.6	34.8
TNIK, TP53	126	124.83	124.83	0.7	0.69	4.65E-02	4.53E-02	35.6	34.8
CDK12, KRAS	105	103.54	103.55	0.86	0.85	4.57E-02	4.46E-02	35.6	34.8
CTCF, KRAS	105	103.55	103.54	0.86	0.86	4.64E-02	4.52E-02	35.6	34.8
DIS3, KRAS	105	103.54	103.56	0.86	0.86	4.45E-02	4.65E-02	35.6	34.8
EGFR, KRAS	105	103.54	103.56	0.85	0.86	4.38E-02	4.60E-02	35.6	34.8
APC, ATM	176	173.97	173.96	1.25	1.23	5.28E-02	4.92E-02	37.9	35.3
APC, FBXW7	179	176.48	176.43	1.76	1.75	7.53E-02	7.08E-02	43.7	44
APC, CTNNB1	175	173.44	173.43	1.07	1.07	7.15E-02	7.14E-02	43.7	44
CEP290, KRAS	104	103.02	103.03	0.71	0.71	8.26E-02	8.55E-02	43.7	44
DNMT3A, KRAS	104	103.03	103.02	0.71	0.71	8.48E-02	8.25E-02	43.7	44
EPN2, KRAS	104	103.02	103.03	0.7	0.71	8.20E-02	8.51E-02	43.7	44
KIFC3, KRAS	104	103.04	103.03	0.7	0.71	8.67E-02	8.43E-02	43.7	44
KRAS, MAP2K1	104	103.01	103.03	0.7	0.71	8.05E-02	8.57E-02	43.7	44
KRAS, NR2F2	104	103.03	103.03	0.7	0.7	8.10E-02	8.31E-02	43.7	44
KRAS, PTPN11	104	103.03	103.03	0.7	0.7	8.37E-02	8.24E-02	43.7	44
KRAS, SF3B1	104	103.02	103.03	0.7	0.7	8.28E-02	8.48E-02	43.7	44
KRAS, STAG2	104	103.03	103.04	0.71	0.7	8.64E-02	8.62E-02	43.7	44
KRAS, TNIK	104	103.04	103.03	0.7	0.71	8.40E-02	8.69E-02	43.7	44
KRAS, WIPF1	104	103.03	103.02	0.71	0.7	8.44E-02	8.23E-02	43.7	44
KRAS, ZC3H11A	104	103.02	103.04	0.7	0.71	8.13E-02	8.73E-02	43.7	44
FBXW7, KRAS	118	114.91	114.9	2.35	2.34	9.42E-02	9.26E-02	46.5	45.7
APC, ARNTL	173	172.36	172.36	0.54	0.55	1.20E-01	1.21E-01	46.7	47.3
APC, CUL1	173	172.36	172.36	0.54	0.54	1.18E-01	1.17E-01	46.7	47.3
APC, EPN2	173	172.35	172.36	0.53	0.54	1.13E-01	1.20E-01	46.7	47.3
APC, GATA3	173	172.36	172.36	0.54	0.55	1.17E-01	1.22E-01	46.7	47.3
APC, LUC7L2	173	172.36	172.35	0.54	0.54	1.18E-01	1.17E-01	46.7	47.3
APC, MAP2K1	173	172.36	172.37	0.54	0.54	1.16E-01	1.23E-01	46.7	47.3
APC, POLR2B	173	172.36	172.35	0.54	0.54	1.17E-01	1.14E-01	46.7	47.3
APC, RUVBL1	173	172.36	172.36	0.54	0.54	1.17E-01	1.19E-01	46.7	47.3
APC, STAG2	173	172.36	172.36	0.54	0.55	1.20E-01	1.23E-01	46.7	47.3
APC, TNIK	173	172.35	172.35	0.54	0.54	1.15E-01	1.14E-01	46.7	47.3
SMAD2, TP53	132	129.79	129.8	1.76	1.79	1.04E-01	1.09E-01	46.7	47.3
FN1, TP53	130	128.12	128.15	1.51	1.51	1.06E-01	1.10E-01	46.7	47.3
CAD, TP53	128	126.49	126.47	1.2	1.19	1.03E-01	9.79E-02	46.7	47.3
APC, DNMT3A	173	172.37	172.37	0.55	0.55	1.23E-01	1.24E-01	46.9	47.3
FBXW7, SMAD4	46	44.33	44.32	1.45	1.46	1.25E-01	1.26E-01	46.9	47.3
AKAP9, APC	174	173.04	173.08	0.91	0.93	1.47E-01	1.60E-01	54.4	57.4
APC, ARID1A	175	173.79	173.81	1.18	1.19	1.53E-01	1.59E-01	55.7	57.4
FBXW7, TCF7L2	44	42.61	42.57	1.38	1.39	1.57E-01	1.52E-01	56.4	56.4
KRAS, NF1	105	104.04	104.06	0.99	0.99	1.67E-01	1.71E-01	58.9	60.3
CDK12, TP53	126	125.24	125.25	0.84	0.85	1.83E-01	1.86E-01	59.7	60.3
CNOT1, TP53	126	125.24	125.25	0.85	0.85	1.85E-01	1.89E-01	59.7	60.3
FOXP1, TP53	126	125.23	125.24	0.85	0.84	1.83E-01	1.85E-01	59.7	60.3
MGA, TP53	126	125.23	125.26	0.85	0.85	1.84E-01	1.91E-01	59.7	60.3
MLL2, TP53	126	125.25	125.24	0.84	0.85	1.88E-01	1.86E-01	59.7	60.3
PIK3R1, TP53	126	125.24	125.23	0.85	0.85	1.85E-01	1.85E-01	59.7	60.3
SMC1A, TP53	126	125.24	125.24	0.84	0.86	1.84E-01	1.87E-01	59.7	60.3
TP53, TRIO	126	125.26	125.24	0.84	0.85	1.89E-01	1.85E-01	59.7	60.3
APC, CASP8	173	172.54	172.54	0.66	0.66	2.45E-01	2.42E-01	69.2	70.3
APC, FOXP1	173	172.53	172.54	0.66	0.67	2.39E-01	2.44E-01	69.2	70.3
APC, GOLGA5	173	172.54	172.55	0.66	0.67	2.42E-01	2.52E-01	69.2	70.3
APC, PIK3R1	173	172.54	172.54	0.66	0.66	2.41E-01	2.45E-01	69.2	70.3
APC, PPP2R1A	173	172.53	172.55	0.66	0.67	2.39E-01	2.50E-01	69.2	70.3
APC, TAF1	173	172.53	172.54	0.66	0.67	2.34E-01	2.44E-01	69.2	70.3
APC, TRIO	173	172.53	172.53	0.66	0.66	2.38E-01	2.40E-01	69.2	70.3
APC, TP53BP1	173	172.53	172.54	0.66	0.66	2.39E-01	2.42E-01	69.2	70.3
CAD, KRAS	106	105.1	105.1	1.22	1.21	2.30E-01	2.29E-01	69.2	70.3
ATM, PIK3CA	44	43.2	43.18	1.19	1.2	2.50E-01	2.47E-01	69.8	70.3
APC, SMAD4	177	175.96	175.96	1.7	1.71	2.71E-01	2.72E-01	74.3	73.7
APC, SOX9	174	173.42	173.44	1.05	1.07	2.92E-01	3.00E-01	74.3	73.7
CTNNB1, TP53	128	127.3	127.33	1.39	1.37	3.07E-01	3.11E-01	74.3	73.7
ARID1A, KRAS	108	107.19	107.15	1.54	1.57	3.00E-01	2.94E-01	74.3	73.7
FN1, KRAS	108	107.18	107.17	1.53	1.54	2.95E-01	2.95E-01	74.3	73.7
CASP8, KRAS	104	103.56	103.55	0.87	0.86	3.07E-01	2.99E-01	74.3	73.7
FOXP1, KRAS	104	103.54	103.55	0.86	0.86	2.97E-01	3.00E-01	74.3	73.7
KRAS, MED12	104	103.52	103.57	0.87	0.86	2.91E-01	3.10E-01	74.3	73.7
KRAS, MGA	104	103.53	103.54	0.86	0.85	2.94E-01	2.96E-01	74.3	73.7
KRAS, PIK3R1	104	103.54	103.54	0.87	0.87	2.99E-01	2.98E-01	74.3	73.7
KRAS, POLR3B	104	103.56	103.56	0.87	0.87	3.05E-01	3.07E-01	74.3	73.7
KRAS, PPP2R1A	104	103.56	103.55	0.86	0.86	3.03E-01	2.98E-01	74.3	73.7
NRAS, PIK3CA	49	48.21	48.2	1.48	1.45	2.97E-01	2.92E-01	74.3	73.7
BRAF, TP53	128	127.34	127.29	1.37	1.38	3.14E-01	3.04E-01	74.9	73.7
GNAS, TP53	128	127.33	127.32	1.39	1.37	3.16E-01	3.08E-01	74.9	73.7
APC, KRAS	191	190.2	190.09	2.65	2.64	3.81E-01	3.65E-01	78.7	79.6
APC, NF1	173	172.72	172.72	0.77	0.76	3.60E-01	3.58E-01	78.7	79.6
FBXW7, TP53	135	134.33	134.36	2.3	2.29	3.86E-01	3.90E-01	78.7	79.6
TCF7L2, TP53	133	132.31	132.28	2.09	2.12	3.70E-01	3.67E-01	78.7	79.6
NF1, TP53	126	125.65	125.64	0.97	0.97	3.61E-01	3.57E-01	78.7	79.6
CDKN1B, TP53	125	124.81	124.83	0.7	0.69	3.95E-01	4.05E-01	78.7	79.6
CEP290, TP53	125	124.82	124.82	0.69	0.7	3.98E-01	3.98E-01	78.7	79.6
CUL1, TP53	125	124.83	124.83	0.7	0.7	4.05E-01	4.02E-01	78.7	79.6
DNMT3A, TP53	125	124.82	124.83	0.69	0.69	3.99E-01	4.03E-01	78.7	79.6
EPN2, TP53	125	124.82	124.83	0.69	0.69	3.98E-01	4.05E-01	78.7	79.6
GATA3, TP53	125	124.83	124.82	0.69	0.69	4.05E-01	3.99E-01	78.7	79.6
MECOM, TP53	125	124.82	124.84	0.7	0.69	4.00E-01	4.06E-01	78.7	79.6
MLLT4, TP53	125	124.83	124.83	0.69	0.7	4.01E-01	4.06E-01	78.7	79.6
POLR2B, TP53	125	124.82	124.82	0.7	0.7	4.01E-01	4.01E-01	78.7	79.6
RUNX1, TP53	125	124.83	124.83	0.7	0.69	4.03E-01	4.02E-01	78.7	79.6
SF3B1, TP53	125	124.83	124.83	0.69	0.69	4.03E-01	4.02E-01	78.7	79.6
TBX3, TP53	125	124.81	124.82	0.69	0.69	3.94E-01	3.99E-01	78.7	79.6
TP53, WT1	125	124.82	124.83	0.69	0.7	3.98E-01	4.03E-01	78.7	79.6
TP53, ZC3H11A	125	124.83	124.82	0.7	0.7	4.02E-01	3.96E-01	78.7	79.6

GENES	COVERAGE	EXPECTATION [N]	EXPECTATION [Nemp]	SD [N]	SD [Nemp]	p [N]	p [Nemp]	% fdr [N]	% fdr [Nemp]
AXIN2, KRAS	106	105.6	105.61	1.3	1.3	3.80E-01	3.82E-01	78.7	79.6
ELF3, KRAS	106	105.62	105.6	1.29	1.3	3.83E-01	3.80E-01	78.7	79.6
PIK3CA, SMAD2	46	45.67	45.71	1.35	1.34	4.03E-01	4.16E-01	78.7	80.8
ATM, KRAS	108	107.68	107.69	1.62	1.61	4.21E-01	4.23E-01	81	81.5
ELF3, TP53	127	126.87	126.87	1.28	1.29	4.61E-01	4.61E-01	88	88.1
APC, NRAS	175	175.04	175.05	1.52	1.51	5.11E-01	5.13E-01	88.6	88.3
APC, RBM10	173	173.05	173.07	0.92	0.93	5.24E-01	5.29E-01	88.6	88.3
SMAD4, TP53	133	133.11	133.13	2.19	2.23	5.20E-01	5.24E-01	88.6	88.3
PCBP1, TP53	126	126.07	126.06	1.09	1.09	5.27E-01	5.21E-01	88.6	88.3
AKAP9, KRAS	105	105.1	105.1	1.21	1.21	5.35E-01	5.33E-01	88.6	88.3
ARNTL, KRAS	103	103.03	103.02	0.71	0.71	5.17E-01	5.14E-01	88.6	88.3
ATRX, KRAS	103	103.05	103.04	0.7	0.7	5.27E-01	5.22E-01	88.6	88.3
CDKN1B, KRAS	103	103.03	103.04	0.7	0.7	5.15E-01	5.23E-01	88.6	88.3
CUL1, KRAS	103	103.02	103.03	0.7	0.71	5.09E-01	5.15E-01	88.6	88.3
KRAS, LUC7L2	103	103.04	103.04	0.71	0.71	5.21E-01	5.21E-01	88.6	88.3
KRAS, MECOM	103	103.02	103.04	0.7	0.7	5.12E-01	5.23E-01	88.6	88.3
KRAS, NUP107	103	103.04	103.04	0.71	0.71	5.21E-01	5.21E-01	88.6	88.3
KRAS, POLR2B	103	103.04	103.02	0.7	0.7	5.22E-01	5.12E-01	88.6	88.3
KRAS, PTEN	103	103.04	103.03	0.71	0.7	5.23E-01	5.18E-01	88.6	88.3
KRAS, RUNX1	103	103.04	103.03	0.71	0.71	5.20E-01	5.15E-01	88.6	88.3
KRAS, TBX3	103	103.02	103.04	0.7	0.7	5.14E-01	5.25E-01	88.6	88.3
KRAS, TCF12	103	103.03	103.01	0.71	0.7	5.17E-01	5.06E-01	88.6	88.3
KRAS, WT1	103	103.03	103.04	0.7	0.7	5.19E-01	5.23E-01	88.6	88.3
CTNNB1, KRAS	106	106.12	106.13	1.39	1.4	5.34E-01	5.37E-01	88.6	88.3
GNAS, KRAS	106	106.13	106.12	1.38	1.41	5.39E-01	5.33E-01	88.7	88.3
KRAS, SMAD2	109	109.22	109.21	1.8	1.82	5.49E-01	5.47E-01	89.7	89.3
APC, TP53	193	193.77	193.74	2.55	2.58	6.19E-01	6.13E-01	94.3	94.3
APC, FN1	173	173.79	173.78	1.17	1.17	7.52E-01	7.48E-01	94.3	94.3
APC, PTPRU	173	173.6	173.6	1.13	1.12	7.02E-01	7.02E-01	94.3	94.3
APC, GNAS	173	173.44	173.43	1.07	1.07	6.59E-01	6.57E-01	94.3	94.3
APC, TGFBR2	173	173.42	173.43	1.05	1.05	6.56E-01	6.58E-01	94.3	94.3
APC, ELF3	173	173.26	173.24	1	1	6.02E-01	5.94E-01	94.3	94.3
APC, CDK12	172	172.54	172.54	0.66	0.66	7.93E-01	7.94E-01	94.3	94.3
APC, CREBBP	172	172.54	172.52	0.66	0.65	7.93E-01	7.88E-01	94.3	94.3
APC, CTCF	172	172.53	172.54	0.66	0.66	7.89E-01	7.90E-01	94.3	94.3
APC, EGFR	172	172.54	172.54	0.66	0.67	7.92E-01	7.93E-01	94.3	94.3
APC, IDH2	172	172.53	172.54	0.66	0.67	7.91E-01	7.93E-01	94.3	94.3
APC, LPHN2	172	172.54	172.53	0.66	0.65	7.91E-01	7.90E-01	94.3	94.3
APC, MED12	172	172.54	172.55	0.66	0.67	7.93E-01	7.96E-01	94.3	94.3
APC, MGA	172	172.52	172.53	0.65	0.66	7.89E-01	7.92E-01	94.3	94.3
APC, MLL2	172	172.54	172.54	0.66	0.66	7.94E-01	7.95E-01	94.3	94.3
APC, POLR3B	172	172.54	172.53	0.66	0.67	7.93E-01	7.88E-01	94.3	94.3
APC, SMC1A	172	172.53	172.54	0.66	0.66	7.89E-01	7.91E-01	94.3	94.3
ACSL6, APC	172	172.36	172.37	0.54	0.54	7.47E-01	7.49E-01	94.3	94.3
APC, ATRX	172	172.36	172.35	0.54	0.54	7.46E-01	7.43E-01	94.3	94.3
APC, CDKN1B	172	172.36	172.36	0.54	0.55	7.48E-01	7.46E-01	94.3	94.3
APC, CEP290	172	172.35	172.37	0.54	0.54	7.45E-01	7.50E-01	94.3	94.3
APC, KIFC3	172	172.36	172.36	0.54	0.54	7.46E-01	7.46E-01	94.3	94.3
APC, MLLT4	172	172.36	172.35	0.54	0.54	7.48E-01	7.42E-01	94.3	94.3
APC, NR2F2	172	172.36	172.36	0.54	0.54	7.45E-01	7.45E-01	94.3	94.3
APC, NUP107	172	172.36	172.35	0.54	0.54	7.47E-01	7.44E-01	94.3	94.3
APC, PTEN	172	172.36	172.36	0.54	0.54	7.47E-01	7.46E-01	94.3	94.3
APC, PTPN11	172	172.35	172.35	0.54	0.54	7.44E-01	7.44E-01	94.3	94.3
APC, RUNX1	172	172.34	172.36	0.53	0.54	7.41E-01	7.50E-01	94.3	94.3
APC, SF3B1	172	172.36	172.37	0.54	0.55	7.44E-01	7.50E-01	94.3	94.3
APC, TBX3	172	172.36	172.36	0.54	0.55	7.46E-01	7.47E-01	94.3	94.3
APC, TCF12	172	172.37	172.35	0.54	0.54	7.52E-01	7.44E-01	94.3	94.3
APC, WIPF1	172	172.36	172.36	0.54	0.54	7.47E-01	7.45E-01	94.3	94.3
APC, WT1	172	172.35	172.36	0.54	0.55	7.45E-01	7.46E-01	94.3	94.3
APC, ZC3H11A	172	172.36	172.37	0.55	0.55	7.45E-01	7.48E-01	94.3	94.3
NRAS, TP53	130	131.03	131.04	1.96	1.95	7.00E-01	7.03E-01	94.3	94.3
PTPRU, TP53	127	127.74	127.72	1.44	1.44	6.95E-01	6.93E-01	94.3	94.3
RBM10, TP53	126	126.46	126.46	1.18	1.2	6.52E-01	6.49E-01	94.3	94.3
DIS3, TP53	125	125.26	125.24	0.84	0.85	6.19E-01	6.11E-01	94.3	94.3
EGFR, TP53	125	125.22	125.24	0.85	0.85	6.04E-01	6.13E-01	94.3	94.3
GOLGA5, TP53	125	125.25	125.25	0.84	0.85	6.15E-01	6.16E-01	94.3	94.3
LPHN2, TP53	125	125.22	125.24	0.85	0.86	6.02E-01	6.12E-01	94.3	94.3
POLR3B, TP53	125	125.24	125.23	0.85	0.83	6.09E-01	6.07E-01	94.3	94.3
TAF1, TP53	125	125.23	125.24	0.84	0.85	6.06E-01	6.09E-01	94.3	94.3
KRAS, SOX9	105	106.14	106.13	1.38	1.4	7.96E-01	7.90E-01	94.3	94.3
CNOT1, KRAS	103	103.55	103.55	0.86	0.86	7.38E-01	7.40E-01	94.3	94.3
CREBBP, KRAS	103	103.54	103.55	0.86	0.86	7.35E-01	7.40E-01	94.3	94.3
GOLGA5, KRAS	103	103.54	103.55	0.87	0.86	7.32E-01	7.37E-01	94.3	94.3
IDH2, KRAS	103	103.54	103.55	0.86	0.86	7.32E-01	7.39E-01	94.3	94.3
KRAS, LPHN2	103	103.55	103.53	0.87	0.86	7.36E-01	7.29E-01	94.3	94.3
KRAS, SMC1A	103	103.55	103.55	0.86	0.87	7.39E-01	7.36E-01	94.3	94.3
KRAS, TAF1	103	103.53	103.54	0.85	0.86	7.33E-01	7.36E-01	94.3	94.3
KRAS, TRIO	103	103.55	103.54	0.86	0.86	7.39E-01	7.37E-01	94.3	94.3
KRAS, TP53BP1	103	103.54	103.55	0.86	0.85	7.34E-01	7.39E-01	94.3	94.3
PIK3CA, TCF7L2	50	50.72	50.76	1.57	1.57	6.77E-01	6.85E-01	94.3	94.3
FAM123B, PIK3CA	46	46.54	46.54	1.38	1.37	6.52E-01	6.53E-01	94.3	94.3
APC, PIK3CA	176	178.1	178.07	2.01	2.04	8.51E-01	8.44E-01	96.5	96.6
APC, TCF7L2	174	175.57	175.58	1.63	1.61	8.33E-01	8.36E-01	96.5	96.6
APC, FAM123B	173	174.7	174.68	1.43	1.42	8.83E-01	8.81E-01	96.5	96.6
APC, AXIN2	172	173.25	173.25	1	0.99	8.95E-01	8.95E-01	96.5	96.6
APC, CAD	172	173.08	173.08	0.93	0.93	8.78E-01	8.77E-01	96.5	96.6
APC, PCBP1	172	172.9	172.91	0.85	0.85	8.55E-01	8.58E-01	96.5	96.6
AXIN2, TP53	125	126.89	126.89	1.29	1.29	9.29E-01	9.29E-01	96.5	96.6
CDC73, TP53	124	125.23	125.25	0.85	0.84	9.27E-01	9.31E-01	96.5	96.6
CTCF, TP53	124	125.23	125.23	0.85	0.85	9.27E-01	9.26E-01	96.5	96.6
IDH2, TP53	124	125.24	125.26	0.85	0.85	9.29E-01	9.31E-01	96.5	96.6
MED12, TP53	124	125.24	125.25	0.86	0.85	9.26E-01	9.29E-01	96.5	96.6
MED24, TP53	124	125.24	125.26	0.85	0.85	9.28E-01	9.29E-01	96.5	96.6
ARNTL, TP53	124	124.82	124.82	0.69	0.7	8.80E-01	8.81E-01	96.5	96.6
KIFC3, TP53	124	124.82	124.82	0.69	0.69	8.83E-01	8.82E-01	96.5	96.6
LUC7L2, TP53	124	124.82	124.83	0.7	0.69	8.80E-01	8.85E-01	96.5	96.6
MAP2K1, TP53	124	124.83	124.83	0.7	0.7	8.83E-01	8.82E-01	96.5	96.6
PTPN11, TP53	124	124.83	124.83	0.7	0.69	8.84E-01	8.83E-01	96.5	96.6
RUVBL1, TP53	124	124.84	124.83	0.7	0.69	8.84E-01	8.84E-01	96.5	96.6
STAG2, TP53	124	124.82	124.82	0.69	0.69	8.83E-01	8.83E-01	96.5	96.6
TP53, WIPF1	124	124.83	124.81	0.7	0.69	8.85E-01	8.81E-01	96.5	96.6
KRAS, SMAD4	110	113.39	113.37	2.22	2.2	9.36E-01	9.37E-01	96.5	96.6
KRAS, TCF7L2	110	112.38	112.32	2.15	2.13	8.67E-01	8.61E-01	96.5	96.6
FAM123B, KRAS	108	109.72	109.73	1.88	1.86	8.20E-01	8.24E-01	96.5	96.6
KRAS, TGFBR2	104	106.11	106.13	1.38	1.4	9.37E-01	9.35E-01	96.5	96.6
KRAS, PCBP1	103	104.57	104.56	1.1	1.1	9.23E-01	9.21E-01	96.5	96.6
ACSL6, KRAS	102	103.04	103.04	0.71	0.71	9.28E-01	9.28E-01	96.5	96.6
GATA3, KRAS	102	103.04	103.03	0.71	0.71	9.28E-01	9.28E-01	96.5	96.6
KRAS, MLLT4	102	103.02	103.03	0.7	0.71	9.27E-01	9.26E-01	96.5	96.6
KRAS, RUVBL1	102	103.03	103.03	0.71	0.7	9.27E-01	9.29E-01	96.5	96.6
FBXW7, PIK3CA	53	54.91	54.9	1.74	1.74	8.64E-01	8.63E-01	96.5	96.6
CDC73, KRAS	102	103.56	103.54	0.85	0.86	9.66E-01	9.63E-01	98.2	98
KRAS, MED24	102	103.56	103.54	0.87	0.87	9.64E-01	9.62E-01	98.2	98
KRAS, MLL2	102	103.56	103.54	0.86	0.87	9.65E-01	9.62E-01	98.2	98
PIK3CA, SMAD4	49	52.39	52.38	1.64	1.66	9.81E-01	9.79E-01	99.3	99.2
KRAS, PIK3CA	111	119.56	119.63	2.63	2.66	9.99E-01	9.99E-01	99.9	99.9
KRAS, PTPRU	103	106.64	106.66	1.45	1.46	9.94E-01	9.94E-01	99.9	99.9
KRAS, RBM10	102	105.09	105.09	1.19	1.19	9.95E-01	9.95E-01	99.9	99.9