

Property-driven statistics of biological networks

Pierre-Yves Bourguignon¹, Vincent Danos², François Képes³, Serge Smidtas¹, and Vincent Schächter¹

¹ Genoscope

² CNRS & Université Paris VII

³ CNRS

Abstract. An analysis of heterogeneous biological networks based on randomizations that preserve the structure of component subgraphs is introduced and applied to the yeast protein-protein interaction and transcriptional regulation network. Shuffling this network, under the constraint that the transcriptional and protein-protein interaction subnetworks are preserved reveals statistically significant properties with potential biological relevance. Within the population of networks which embed the same two original component networks, the real one exhibits simultaneously higher bi-connectivity (the number of pairs of nodes which are connected using both subnetworks), and higher distances. Moreover, using restricted forms of shuffling that preserve the interface between component networks, we show that these two properties are independent: restricted shuffles tend to be more compact, yet do not lose any bi-connectivity.

Finally, we propose an interpretation of the above properties in terms of the signalling capabilities of the underlying network.

1 Introduction

The availability of genome-scale metabolic, protein-protein interaction and regulatory networks [25, 7, 3, 5, 21] —following closely the availability of large graphs derived from the Internet hardware and software network structure, from social or collaborative relationships— has spurred considerable interest in the empirical study of the statistical properties of these ‘real-world’ networks. As part of a wider effort to reverse-engineer biological networks, recent studies have focused on identifying *salient* graph properties that can be interpreted as ‘traces’ of underlying biological mechanisms, shedding light either on their dynamics [23, 11, 6, 28] (*i.e.*, how the connectivity structure of the biological process reflects its dynamics), on their evolution [10, 30, 27] (*i.e.*, likely scenarios for the evolution of a network exhibiting the observed property or properties), or both [9, 14, 15]. The statistical graph properties that have been studied in this context include the distribution of vertex degrees [10, 9], the distribution of the clustering coefficient and other notions of density [17–19, 22, 4], the distribution of vertex-vertex distances [22], and more recently the distribution of network motifs occurrences [15].

Identification of a salient property in an empirical graph —for example the fact that the graph exhibits a unexpectedly skewed vertex degree distribution— requires a prior notion of the distribution of that property in a class of graphs relatively to which saliency is determined. The approach chosen by most authors so far has been to use a *random graph model*, typically given by a probabilistic graph generation algorithm that constructs graphs by local addition of vertices and edges [20, 1, 24]. For the simplest random graph models, such as the classical Erdős-Rényi model (where each pair of vertices is connected with constant probability p , [2]), analytical derivations of the simplest of the above graph properties are known [20, 1].

In the general case, however, analytical derivation is beyond the reach of current mathematical knowledge and one has to retort to numerical simulation. The random graph model is used

to generate a sample of the corresponding class of graphs and the distribution of the graph property of interest is evaluated on that sample, providing a standard against which the bias of the studied graph can be measured [23, 14, 29]. Perhaps because of the local nature of the random graph generation process, it is mostly simple *local* network properties that have been successfully reproduced in that fashion. Another, somewhat more empirical, category of approaches reverses the process: variants are generated from the network of interest using a random rewiring procedure. The procedure selects and moves edges randomly, preserving the global number of edges, and optionally their type, as well as local properties such as the degree of each vertex. Rewirings are thus heuristic procedures which perform a sequence of local modifications on the structure of the network.

The specific focus of the present paper is on measuring the degree of cooperation between the two subgraphs of the yeast graph of interactions induced by the natural partition of edges as corresponding either to transcriptional interaction (directed) or to protein protein interaction (undirected). To evaluate a potential deviation with respect to such a measure, one needs as a first ingredient a suitable notion of random variation of the original graph. The goal is here, as in many other cases, to contrast values of a given observable on the real graph, against the distribution of those same values in the population of variants. We define *shuffles* of the original graph as those graphs that are composed exactly of the original two subgraphs of interest, the variable part being the way these are ‘glued’ together.

From the probabilistic point of view, this notion of randomisation coincides with a traditional Erdős-Renyi statistics, except that it is conditioned by the preservation of the original subgraphs. Designing a generative random graph model that would only yield networks preserving this very precise property seems to be a hard endeavor ; it is not as easy as in the unconditional Erdős-Renyi model to draw edges step by step yet ensure that component subgraphs will be obtained in the end. Shuffling might also be seen as rewiring, except the invariant is large-scale and extremely precise: it is not edges that are moved around but entire subgraphs. Moving edges independently would break the structure of the subnetworks, and designing a sequential rewiring procedure that eventually recovers that structure is not an obvious task. Moreover, it would be in general difficult to ensure the uniformity of the sample ; see [16] for a thorough analysis of rewiring procedures. This choice of an invariant seems rather natural in that one is interested in qualifying the interplay between the original subgraphs in the original graph. Now, it is not enough to have a sensible notion of randomisation, it is also crucial to have a computational handle on it. Indeed, whatever the observable one wants to use to mark cooperation is, there is little hope of obtaining an analytic expression for its distribution, hence one needs sampling. Fortunately, it turns out it is easy to generate shuffles uniformly, since these can be described by pairs of permutations over nodes, so that one can always sample this distribution for want of an exact expression. As explained below in more details, the analysis will use two different notions of subgraph-preserving sampling: *general* shuffles, and *equatorial* ones that also preserve the interface between our two subgraphs. Equatorial shuffles are feasible as well, and in both cases the algorithms for sampling and evaluating our measures turn out to be fast enough so that one can sweep over a not so small subset of the total population of samples.

Regarding the second necessary ingredient, namely which observable to use to measure in a meaningful way the otherwise quite vague notion of cooperation, there are again various possibilities. We use two such observables in the present study: the *connectivity*, defined as the percentage of disconnected pairs of nodes, and a refined quantitative version of connectivity, namely the full distance distribution between pairs of nodes. The latter is costlier, requiring about three hours of computation for each sample on a standard personal computer.

Once we have both our notion of randomisations and our observables in place, together with a feasible way of sampling the distribution of the latter, we can start. Specifically we run four

experiments, using general or equatorial shuffling, and crude or refined connectivity measures. The sampling process allows us to compare the values of these measures for the original graph with the mean value for the sample, and, based on the assumption that those values follow a normal distribution over the sample, one can also provide a p -value that gives a rough estimate of the statistical deviation of the observable in the given graph.

The general shuffle based experiments show with significant statistical confidence that shuffling reduces connectivity (1), and at the same time contracts distances (2). More precisely, both bi-connectivity (the amount of pairs of nodes which are connected using both subgraphs) and distances are higher than average in the real network. A first interpretation might be that the real graph is trading off compactness for better bi-connectivity. In order to obtain a clearer picture and test this interpretation, we perform two other experiments using equatorial shuffles. Surprisingly, under equatorial shuffles connectivity hardly changes, while the global shift to shorter distances is still manifest. It seems therefore there is actually no trade-off, and both properties (1) and (2) have to be thought of as being independently captured by the real graph. With appropriate caution, we may try to provide a biological interpretation of this phenomenon. Since all notions of connectivity and distances are understood as directed, we propose to relate this to signalling, and interpret bi-connectivity as a measure of the capability to convey a signal between subgraphs. With this interpretation, the above properties may be read as: (1) signal flows better than average and (2) signal is more specific than average. The second point requires explanation. At constant bi-connectivity, longer average distances imply that upon receipt of a signal, the receiver has a better chance of guessing the emitter. In other words, contraction of distances (which can be easily achieved by using hubs) will anonymise signals, clearly not a desirable feature in a regulatory network. Of course this is only part of the story, since some hubs will also have an active role in signal integration and decision making. The latter is probably an incentive for compactness. If our reading of the results is on track, we then may think of the above experiments as showing that the tropism to compactness due to the need for signal integration, is weaker than the one needed for signal specificity.

Beyond the particular example we chose to develop here because of the wealth of knowledge available on the yeast regulatory and protein interaction networks, one can think of many other applications of the shuffling methodology for heterogeneous networks. The analyses performed here rely on edges corresponding to different types of experimental measurements, but edges could also represent different types of predicted functional links. Indeed, there are many situations where a biological network of interactions can be naturally seen as heterogeneous. Besides, the notions of shuffle we propose can also accommodate the case where one would use a partition of nodes, perhaps given by clustering, or localisation, or indeed any relevant biological information, and they may therefore prove useful in other scenarios.

The paper is organised as follows: first, we set up the definitions of edge-based general and equatorial shuffles based, and also consider briefly node-based shuffles though these are not used in the sequel; then we describe the interaction network of interest and the way it was obtained; finally we define our observables and experiments, and interpret them. In the conclusion, we discuss generalization and potential applications of the method. The paper ends with an appendix on the algorithmical aspects of the experiments, and a brief recall of the elementary notions of statistics we use to assert their significance.

2 Shuffles

Let $G = (V, E)$ be a directed graph, where V is a finite set of nodes, and E is a finite set of directed edges over V . We write M for the incidence matrix associated to G . Since G is directed,

M may not be symmetric. In the absence of parallel edges M has coefficients in $\{0, 1\}$, where parallel edges are allowed.

Given such a matrix M and a permutation σ over V , one writes $M\sigma$ for the matrix defined as for all u, v in V :

$$M\sigma(u, v) := M(\sigma^{-1}u, \sigma^{-1}v)$$

Note that $M\sigma$ defines the same abstract graph as M does, since all σ does is changing the nodes names.

2.1 Shuffles Induced by Properties on Edges

We consider first shuffles induced by properties on edges. Suppose given a partition of $E = \sum E_i$; this is equivalent to giving a map $\kappa : E \rightarrow \{1, \dots, p\}$ which one can think of as colouring edges.

Define M_i as the incidence matrix over V containing the edges in E_i (of colour i).

Define also V_I , where $I \subseteq \{1, \dots, p\}$, as the subset of nodes v having for each $i \in I$ at least one edge incident to v with colour i , and no incident edge coloured j for $j \notin I$. We abuse notation and still write $\kappa(u) = I$ when $u \in V_I$. This represents the set of colours seen by the nodes.

Clearly $V = \sum V_I$, V_\emptyset is the set of isolated nodes of G , and the set of nodes of G_i is the union of the graphs generated by V_I , for $i \in I$.

Given $\sigma_1, \dots, \sigma_p$ permutations over V , define the *global shuffle* of M as:

$$M(\sigma_1, \dots, \sigma_p) := \sum_i M_i \sigma_i$$

The preceding definition of $M\sigma$ is the particular case where $p = 1$ (one has only one colour common to all edges). Each G_i (the abstract graph associated to M_i) is preserved up to isomorphism under this transformation. However the way the G_i s are glued together is not, since one uses a different local shuffle on each.

For moral comfort, we can check that any means of glueing together the G_i s is obtainable using a general shuffle in the following sense: given G' and $\sum q_i : \sum G_i \rightarrow G'$ where the disjoint sum $\sum_i q_i$ is an isomorphism on edges, one has that G' is a general shuffle of G . To see this, define $\sigma_i(u) := q_i p_i^{-1}(u)$ if $u \in \kappa^{-1}(i)$, $\sigma_i(u) = u$ else (we have written p_i for the inclusion of G_i in G), one then has $G' = \sum G_i \sigma_i = G(\sigma_1, \dots, \sigma_p)$.

Note also that $(M(\sigma_1, \dots, \sigma_p))\tau := \sum_i M_i(\tau \sigma_i)$, and so in particular, without loss of generality one can take any the σ_i 's to be the identity (just take $\tau = \sigma_i^{-1}$). This is useful when doing actual computations, and avoids some redundancy in generating samples.

An additional definition will help us refine the typology of shuffles. One says a shuffle $M\sigma$ is *equatorial* if in addition for all I , and all i , V_I is closed under σ_i . Equivalently, one can ask that $\kappa \circ \sigma_i = \kappa$. An *equatorial shuffle* preserves the set of colours associated with each node and in particular preserves for a given pair of nodes (u, v) the fact that (u, v) is heterochromatic, *i.e.*, $\kappa(u) \cap \kappa(v) = \emptyset$. This in turn implies that the distance between u and v must be realised by a path which uses edges of different colours. In the application such paths are mixing different types of interaction, and are therefore of particular interest; without preserving this attribute, an observable based on path with different colours would not make sense. In the particular case of two colours, nodes at the 'equator', having both colours, will be globally preserved, hence the name.

2.2 Shuffles Induced by Properties on Vertices

One can also consider briefly shuffles induced by properties on nodes. Suppose then given a partition of nodes $V = \sum_i V_i$, again that can be thought of as a colouring of nodes $\kappa : V \rightarrow \{1, \dots, p\}$, and extended naturally to the assignment of one or two colours to each edge.

A node shuffle is defined as a shuffle associated to σ which can be decomposed as $\sum_i \sigma_i$, σ_i being a permutation over each cluster V_i . Clearly each graph G_i generated by V_i is invariant under the transformation: only the inter-cluster connectivity is modified.

The equivalent of the equatorial constraint would be to require in addition $\sigma(u) \in \partial V_i$ if $u \in \partial V_i$, where ∂V_i is defined as those nodes of V_i with an edge to some V_j , $i \neq j$. Other variants are possible and the choice of the specific variant will likely depend on the particular case study. We now turn to the description of the network the shuffle experiments will be applied to.

3 A Combined Network of Regulatory and Protein-Protein Interactions in Yeast

With our definitions in place, we can now illustrate the approach on a heterogeneous network obtained by glueing two component networks.

It is known that regulatory influences, including those inferred from expression data analysis or genetic experiments, are implemented by the cell through a combination of direct regulatory interactions and protein-protein interactions, which propagate signals and modulate the activity level of transcription factors. The detailed principles underlying that implementation are not well understood, but one guiding property is the fact that protein interaction and transcriptional regulation events take place in the regulatory network at different time-scales.

In order to clarify the interplay between these two types of interactions, we have combined protein-protein (PPI) and protein-DNA (TRI, for ‘transcriptional regulation interaction’) interaction data coming from various sources into a heterogeneous network by glueing together these two networks on the underlying set of yeast proteins.

The data from which the composite network was built includes: 1440 protein complexes identified from the literature, through HMS-PCI or TAP [3, 5], 8531 physical interactions generated using high-throughput Y2H assays [26], and 7455 direct regulatory interactions compiled from literature and from CHIP-Chip experiments [4, 12], connecting a total of 6541 yeast proteins. A subnetwork of high-reliability interactions was selected, using a threshold on the confidence levels associated to each inferred interaction. For the CHIP-Chip data produced by Lee et al. [12], interactions with a p -value inferior to 3.10^{-2} were conserved ; for the Y2H data produced by Ito et al. [26], a threshold of 4.5 on the Interaction Sequence Tag was used (see [8]). The PPI network was built by connecting two proteins, in both directions, whenever there was a protein-protein or a complex interaction between the two corresponding proteins. In the case of the TRI network, an edge connects a regulator protein with its regulatee. To simplify the discussion, we will refer in the rest of the paper to the TRI graph as TRI , and to the PPI graph as PPI . With some more precision, define G as the real graph, TRI as the subgraph induced by the set of TRI nodes, *i.e.*, nodes such that $TRI \in \kappa(u)$, and PPI as the subgraph induced by the set of PPI nodes.

Their respective sizes are:

$$TRI = 3387, PPI = 2517, TRI \cup PPI = 4489, TRI \cap PPI = 1415$$

The set of nodes $TRI \cap PPI$ of both colours is also referred to in the sequel as the *equator* or the *interface*. Since the object of the following is to discuss the interplay between the TRI and PPI subgraphs, the interface naturally plays an important role. A qualitative measure of the connectivity between TRI and PPI which will be useful later in the discussion, is the number of bi-connected pairs in G (these are the pairs which are connected in G , but not connected in either TRI or PPI), which is roughly $p_{bi} = 23\%$. To complete this statistical portrait of the data, we provide in figure 1 the histograms of degree distributions in the PPI and TRI networks, with in and out degrees pictured separately for the latter. Figure 1 also shows the hub size distribution

for the TRI network (the PPI network has no non-trivial hubs). Note that hubs are defined as sets of nodes connected to a single node. The TRI network (here considered as unoriented) has 124 such hubs ; the histogram of the distribution of their sizes is given in figure 1.

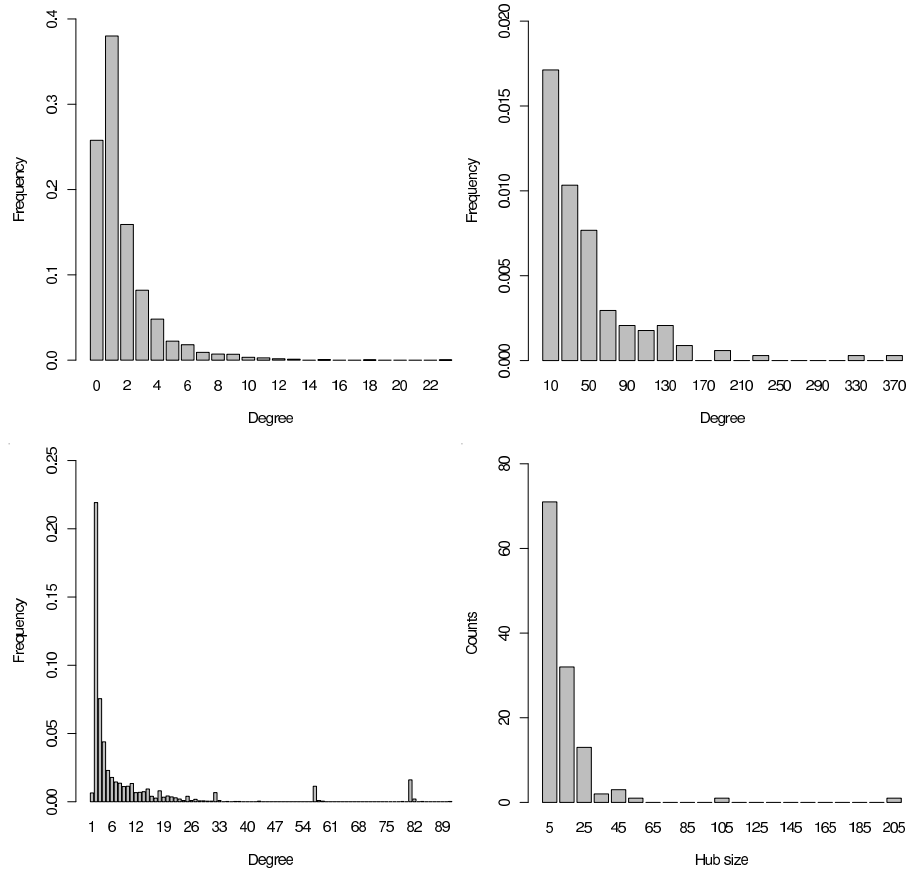


Fig. 1. First row: Histograms of the in and out degree distributions of the TRI network. Second row: Histogram of the degree distribution of the PPI network and of the distribution of the hub size in the TRI network.

4 Results and Interpretations

Hereafter, notions of connectivity, distance, etc. should be understood as *directed* unless explicitly stated otherwise. We now turn to the various shuffle experiments and consecutive observations.

4.1 General Shuffle vs Connectivity

We take here as a rough measure of the connectivity of a graph the percentage of unconnected pairs. Comparing first the real graph with the randomised versions under the general shuffle, one

finds that in the average 4% of the population pairs are disconnected under shuffle. So general shuffle disconnects, or in other words G maximises bi-connectivity.

Clearly mono-connected pairs (pairs connected in either PPI or TRI) cannot be disconnected under general shuffle; a pair is ‘breakable’ only if bi-connected in G ; therefore a more accurate measure of the connectivity loss under general shuffle is that about 17.5% of the breakable pairs are actually broken (this obtained by dividing by p_{bi}), a rather strong deviation with a p -value below 10^{-11} .

Inasmuch as a directed path can be thought of as a signal-carrying pathway, one can interpret the above as saying that the real graph connects PPI and TRI so as to maximise the bandwidth between the subgraphs.

4.2 Equatorial Shuffle vs Connectivity

Keeping with the same observable, we now restrict to equatorial shuffles. One sees in this case that no disconnection happens, and actually about 1% more pairs are connected *after* shuffling. The default of connected pairs of the real graph has a far less significant p -value of 3%. However the point is that equatorial shuffles leaves bi-connectivity rather the same.

This complements the first observation and essentially says that the connectivity maximisation seen above is a property of the set of equatorial nodes ({TRI,PPI} nodes) itself, and not of the precise way TRI and PPI edges meet at the equator.

Both observations can be understood as saying that the restriction of G to the equator is a much denser subgraph than its complement (as evidenced by the connectivity loss under general shuffle), and dense enough so that equatorial shuffling does not impact connectivity.

Note that so far the observable is somewhat qualitative, being only about whether a pair is connected or not. Using a refined and quantitative version of connectivity, namely the distribution of distances (meaning for each n the proportion of pairs at distance n), will reveal more.

4.3 Impact of Equatorial Shuffles on Distance Distribution

Using this refined observable, one sees that the whole histogram shifts to the left, so equatorial shuffle contracts the graph (Fig. 2). This is confirmed by the equality between the number of lost pairs at distance 7 to 9 and the number of new ones at distance 3 to 5. In accordance with the preceding experiment, one also does not see any disconnection under equatorial shuffle.

This is to be compared with the general shuffle version (Fig. 3) where both effects are mixed, and the cumulated excess of short pairs does not account for the loss of long pairs (indeed we know 4% are broken, *i.e.*, disappear at infinity and are not shown on the histogram).

To summarize the distance distribution results in a single number, one can compute the deviation of the real graph mean distance under both shuffles. As expected the mean distance is higher in the real graph with respective p -values of 0.2% and 2% in the general and equatorial shuffles (see Appendix for details). We conclude that while the real graph does maximise bi-connectivity, it does not try to minimise the associated distances.

To provide an intuition on the potential interpretation of the above result, let us again consider paths as rough approximations of signalling pathways. Now compare a completely linear chain-shaped graph and star-shaped one, with the same number of nodes and edges. In the star case, any two nodes are close, at constant distance 2, while in the chain distances are longer. As said, compactness comes with a price, namely that in a star graph all signals go through the hub and are anonymised, *i.e.*, there may be a signal, but there is no information whatsoever in the signal about where the signal originated from. Quite the opposite happens in a linear graph. Of course this is an idealized version of the real situation; nevertheless it is tempting to interpret this last

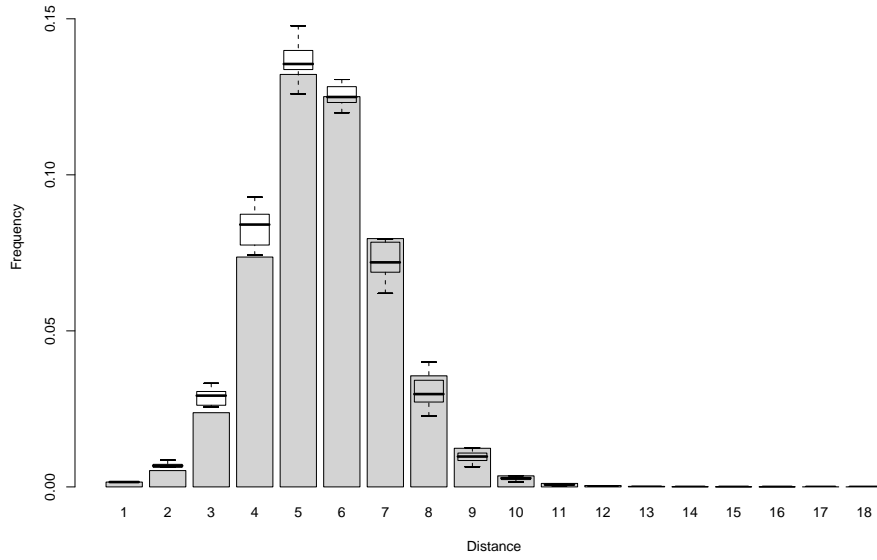


Fig. 2. Equatorial shuffle distance histogram: grey boxes stand for the real graph; one sees that shuffles have more pairs at shorter distance, and consequently (because the number of connected pairs is about the same) less such at higher distances.

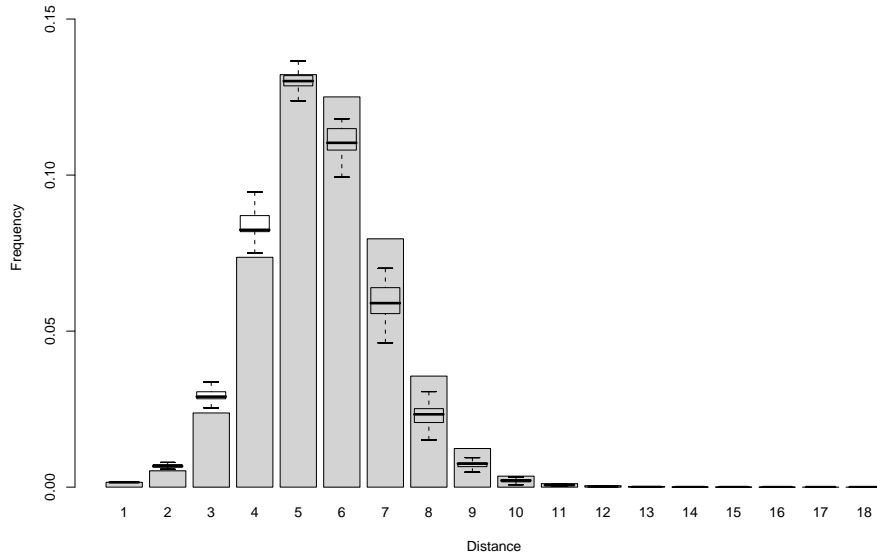


Fig. 3. Global shuffle distance histogram

observation as an indication that the real graph is trading off fast connectivity against specificity of signals. The heterogeneous network is likely to result from a trade-off between causality and signal integration.

As suggested in the introduction, finer observables would have to be developed to further refine this interpretation. Furthermore, there are intrinsic limits on the nature of properties that can be identified using pure topology; deeper, reliable insights about signal transmission in the joint network will ultimately require a dynamical view of signaling with corresponding experimental data.

5 Conclusion

In order to assess the cooperation between the network of protein-protein interactions and the regulatory network in yeast, we have defined two notions of shuffle, *i.e.* tractable randomisations of the original network that preserve global invariants. While general shuffles preserve the entire structure of the component subnetworks, equatorial shuffles also preserve the interface between the networks. We assessed cooperation between the subnetworks using two observables: the percentage of connected node pairs, and the distribution of distances between nodes. For each shuffle-observable pair, the observable in the real network was assessed against the distribution of observables in the set of network variants generated by the respective shuffle.

To summarise the results of this case study, we can say that the statistical analysis of G shuffles under the constraint of preserving its component subnetworks suggests the existence of two *independent* properties of G regarding the cooperation of its components:

- bi-connectivity, *i.e.* the proportion of node pairs connected only by paths using both types of edges, is higher in the actual network than in the shuffles;
- distances between pairs of nodes are higher in the actual network;

The first property can be given an interpretation in terms of bandwidth: signals flow better between the two networks than would be expected if they were connected randomly. The second property can be interpreted as favoring signal specificity: for cellular interaction networks (in contrast with telecommunication networks, for instance, where each packet carries significant intrinsic information) the information borne by a signal is very much related to the path it has followed. Longer paths thus provide more opportunity for specific signals. Note that the fact that we worked with directed notions (and not with undirected ones as we did in a first version of this paper) makes the interpretation of paths as potential signaling pathways somewhat more convincing.

We have been careful in the discussion of the results of our statistical experiments in terms of signalling capacities, and this needs to be thrashed out in subsequent work. To do so one would first need refined and yet feasible observables pertaining to the dynamics of the network of interest. A recent paper equips the subgraph induced by the major molecular players in the budding yeast cell cycle (cyclins, their inhibitors, and major complexes) with a discrete Boolean dynamics [13], and obtains a dynamics with a stable state corresponding to the G_1 phase, which is attracting a significantly higher number of states than a random graph (with the same number of nodes and edges) would. It seems therefore possible to explicitly construct signal-related observables. However there are several problems: first, this analysis relies on sorting positive and negative regulation edges, and that is an information which one doesn't have for the full graph; second it also critically relies on the rather small size of the subgraph; finally the model only handles a limited number of signals (corresponding to the various cell cycle phases). Nevertheless, a comparable study, using shuffles as a means of randomising, and confined to a

well-chosen subgraph could help in qualifying our speculative interpretation of the contraction phenomenon we have observed.

On the methodological front, both the general notion of shuffle and the restricted notion of equatorial shuffle proved useful: they reveal different properties and complement one another. The same holds for the pair of observables: both the qualitative connectivity observable and its refined distance-based version are useful, and yield different and complementary insights on the cooperation between the two component networks.

We believe that the shuffling methodology developed for this case study has general applicability to the study of heterogeneous biological networks, i.e. networks that can be seen as the “glueing” of two or more component networks. Shuffles preserve global invariant properties (the structure of component networks), and define rigorously and unambiguously the class of networks which obey these properties. They are also easily computable and can be generated uniformly, by drawing from a set of acceptable permutations. Note that the latter property is in contrast with randomizations based on sequential rewiring strategies, where each rewiring step perturbs the structure while preserving one or more local invariants. While these approaches may prove to be asymptotically equivalent in some cases, they typically do not provide a direct definition nor the means to uniformly sample the set of randomizations which preserve the invariant, since the order of the rewiring steps matters.

Given an interaction network between biochemical species, any biological property on edges (type of interaction, degree of confidence, localization of interaction...) or on nodes (type of entity, functional annotation, inclusion within clusters generated using a given data type and methodology, etc...) with discrete values can be used to define a heterogeneous version of that network. Then, either the type of edge shuffles used above, or shuffles preserving other categories of top-down invariants, such as the projection of a network onto a given network of abstract clusters, could be explored. Likewise, a variety of observable properties may be used to investigate cooperation between component subnetworks. Perhaps the foremost promise of the shuffling approach resides in the interplay between different shuffle-observable pairs, which allows an exploratory assessment of cooperation adapted to the heterogeneous network at hand.

References

1. William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In *32nd Annual ACM Symposium on Theory of Computing*, pages 171–180, 2000.
2. P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:1761, 1960.
3. AC Gavin, M Bosche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, JM Rick, AM Michon, CM Cruciat, M Remor, C Hofert, M Schelder, M Brajenovic, H Ruffner, A Merino, K Klein, M Hudak, D Dickson, T Rudi, V Gnau, A Bauch, S Bastuck, Huhse, C Leutwein, MA Heurtier, RR Copley, A Edelmann, E Querfurth, V Rybin, G Drewes, M Raida, T Bouwmeester, P Bork, B Seraphin, B Kuster, G Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868)(Jan 10):141–7., 2002.
4. N. Guelzim, S. Bottani, P. Bourguin, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–3, 2002.
5. Y Ho, A Gruhler, A Heilbut, GD Bader, L Moore, SL Adams, A Millar, P Taylor, K Bennett, K Boutilier, L Yang, C Wolting, I Donaldson, S Schandorff, J Shewnarane, M Vo, J Taggart, M Goudreault, B Muskat, C Alfarano, D Dewar, Z Lin, K Michalickova, AR Willems, H Sassi, PA Nielsen, KJ Rasmussen, JR Andersen, LE Johansen, LH Hansen, H Jespersen, A Podtelejnikov, E Nielsen, J Crawford, V Poulsen, BD Sorensen, J Matthiesen, RC Hendrickson, F Gleeson, T Pawson, MF Moran, D Durocher, M Mann, CW Hogue, D Figeys, and M Tyers. Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868)(Jan 10):180–3, 2002.

6. J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31(4):370–7, 2002.
7. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–74, 2001.
8. Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Masahira Yoshida, Mikio and Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS*, 98(8):4569–4574, 2001.
9. H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, 2001.
10. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.
11. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 2004.
12. T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002.
13. Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101(14):11250–11255, April 2004.
14. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–42, 2004.
15. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–7, 2002.
16. R. Milo, S. S. Shen-Orr, S. Itzkowitz, N. Kashtan, D. Chklovskii, and U. Alon. On the uniform generation of random graphs with prescribed degree sequence. *ArXiv*, 2003.
17. M. E. Newman. Assortative mixing in networks. *Phys Rev Lett*, 89(20):208701, 2002.
18. M. E. Newman. Properties of highly clustered networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(2 Pt 2):026121, 2003.
19. M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2):026113, 2004.
20. M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(2 Pt 2):026118, 2001.
21. N. D. Price, J. A. Papin, C. H. Schilling, and B. O. Palsson. Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol*, 21(4):162–9, 2003.
22. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, 2002.
23. S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *escherichia coli*. *Nat Genet*, 31(1):64–8, 2002.
24. S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–76, 2001.
25. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770):623–7, 2000.
26. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.
27. A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18(7):1283–92, 2001.
28. D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, 1998.
29. E. Yeager-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, 101(16):5934–9, 2004.
30. S. H. Yook, H. Jeong, A. L. Barabasi, and Y. Tu. Weighted evolving networks. *Phys Rev Lett*, 86(25):5835–8, 2001.

A Computation of the Shortest Paths Length Distribution

This section is devoted to a brief description of the algorithms and methods used to derive the various statistics used in the study of the yeast regulation and protein interaction networks.

Clearly the (i, j) coefficient of M^n is the number of oriented paths of length n connecting i to j in the graph underlying M . Since we are only interested in knowing whether two nodes are connected by an oriented path of a given length we may use a simplified matrix product defined as:

$$M^n(i, j) = \begin{cases} 1 & \text{if } \exists k \in V : M^{n-1}(i, k) = M(k, j) = 1 \\ 0 & \text{otherwise} \end{cases}$$

which is just forgetting the numbers of connecting paths, only to remember whether there is at least one.

Furthermore, the addition of the identity matrix I to the adjacency matrix before the computation of the products gives an immediate access to the value of the cumulative distribution function of the oriented, shortest path length distances in the network. Indeed, writing $\widehat{M} = M + I$:

$$\widehat{M}^n(i, j) = \begin{cases} 1 & \text{if } \exists k \in V, M^{n-1}(i, k) = M(k, j) = 1 \\ & \text{or } \widehat{M}^{n-1}(i, j) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus the number of 1s in $\widehat{M}(G)^n$ is the number of ordered pairs connected by at least one path of length $\leq n$, and the entire distribution is obtained when the computation reaches a fixpoint. Computing the distribution on the real PPI-TRI graph takes about 180' on a recent PC ; the distribution for the 100 shuffles were computed down in less than 10 hours on a cluster of 41 computers hosted by Genoscope.

B Statistics

This section details the definition and computation of p -values shown in the statistical results, concerning both the amount of connected pairs and the average distance.

In order to compute p -values for the deviation of the observable on the real graph from its distribution over the set of shuffled ones, we need to approximate this distribution by a Gaussian one, with mean and standard deviation fixed to the empirical values computed on the sample. This is necessary, since the rather low amount of shuffled networks (100) prevents a direct estimation of the p -value as the proportion of shuffled networks with a larger observable.

Concerning the amount of disconnected pairs, which is the first observable considered in the results, the empirical mean over the set of general shuffles is $m_g = 0.574$, and the standard deviation $s_g = 0.005$. In the case of the equatorial shuffle, the mean falls to $m_e = 0.534$, with a standard deviation of 0.002.

Assuming this average proportion is a Gaussian random variable A with those parameters, the p -value of the deviation of the average proportion of disconnected pairs in the real network from its distribution over the sample of general shuffled networks is defined as:

$$p_g = \mathbb{P}(A < m_G), \quad \text{with } A \sim \mathcal{N}(m_g, s_g)$$

where $m_G = 0.538$ is the observed proportion of disconnected pairs in the real network. In this case, this yields $p_g = 9 \times 10^{-12}$.

Since the proportion of disconnected pairs in the real graph is higher than the average amount of disconnected pairs in the equatorially shuffled ones, one computes the p -value p_e using the upper tail of the distribution instead of the lower one:

$$p_e = \mathbb{P}(A > m_G), \quad \text{with } A \sim \mathcal{N}(m_e, s_e)$$

so that $p_e = 0.03$.

The computation of the p -value for the deviation of the mean distance from its value on shuffled networks follows the same scheme. The mean distance in the real graph is $m_G^d = 5.66$, while its average over the set of shuffled graphs is $m_g^d = 5.38$ for the general shuffle, and $m_e^d = 5.5$ for the equatorial one. Standard deviation is $s_g^d = 0.09$ with the general shuffle, and $s_e^d = 0.08$ with the equatorial one. The p -values for these deviations are $p_g^d = 0.002$ and $p_e^d = 0.02$, respectively.