1 Freie Universität Berlin

2 International Max Planck Research School for Computational Biology and Scientific Computing

3 Berlin Mathematical School

4 DFG Research Center MATHEON Mathematics for key technologies

5 nanopoly

Understanding biological networks with the random walker's perspective

Sharon Bruckner[1,2,5] Tim OF Conrad[1,4], Natasa Djurdjevac[1,3], Christof Schütte[1,4]

Introduction

Background: Studying the organization of modular biological networks such as protein-protein interaction networks (PPI) can bring insights into the dynamics of the processes in the cell.

Aim: reveal the organization of biological networks. While previous methods aim to identify a specific type of network element (community, hub, etc.), our methods discover these individual elements as well as the connections between them, detecting modules, identifying the important paths between them, and pinpointing **key nodes** in a network, which are most vital to network communication.

Methods: We base our novel algorithms on the well-known *random walker approach*: we equate modules with metastable sets. We then associate the important paths and key nodes with those carrying the most flow in the sense of *Transition Path Theory (TPT)*, a rigorous framework with proven properties originally designed for the study of dynamical systems.

Results: We demonstrate the effectiveness of our methods in recovering known structures from a yeast PPI network.

Methods

A common approach for the analysis of networks exploits the strong connection between networks and Markov chains. One goal is to connect dynamical properties of the associated random walk to structural properties of the network itself. We then equate modules with node sets that will be metastable in the sense of the associated Markov chain: regions of the network where the random walker stays for a long time before exiting.

(1) Constructing the generator L.

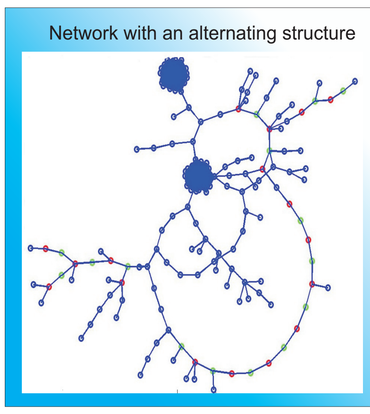
One drawback of using the standard random walker with transition matrix P to detect dense network substructures is that structures like long chains are also metastable: The chain is decomposed into two sets, and the random walker alternates between them. We consider instead a time-continuous Markov process with generator L having the form of a rate matrix defined a

$$L(x,y) = \begin{cases} -\frac{1}{d(x)^p}, & x = y \\ \frac{1}{d(x)^{p+1}}, & x \neq y, (x,y) \in E \\ 0, & else. \end{cases}$$

where $d(x)$ is the degree of node x.
Unlike the discrete random walker, that moves from node to node in every timestep, the **continuous** walker must wait for the jump for a specified *waiting time*. The waiting time $d(x)^p$ that emerges from the definition above is proportional to the degree of the node: the more neighbors the random walker has to choose from, the longer the decision time. Thus, chains are no longer problematic, as the random walker moves along them quickly with **low** waiting time, while dense structures still slow it down with a **high** waiting time and are metastable.

Relevant paper: *Djurdjevac, N., Sarich, M., Bruckner, S., Conrad, T., Schuette, Ch. Manuscript in preparation.*

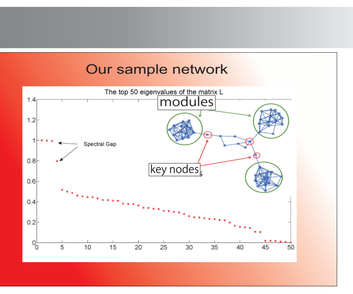
Network with an alternating structure



(2) Establishing the number of modules.

We first determine the number of modules in the network.
This we obtain by looking at the spectrum of the transition matrix P' generated by L by $P' = exp(L*t)$
and counting the number of dominant eigenvalues: those eigenvalues close to 1 that control the behavior of the random walker in the long term. The **spectral gap** is clear due to the nice properties of P' (for example, P' never has negative eigenvalues, unlike the standard transition matrix) and we can easily estimate the number of modules.

Our sample network



(3) Identifying the modules.

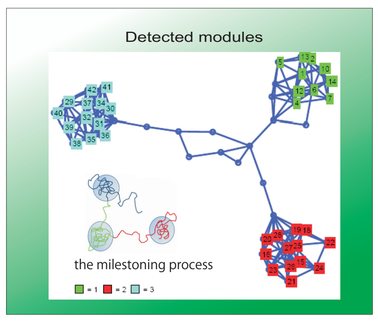
We are not looking for a full partition of the network, rather we wish to **milestone** the network so that the behavior of the random walker on the milestoned network is the same as on the original network. That is done by looking for modules such that the spectrum of the milestoning process is close to that of the original process. More exactly, we wish to minimize the **eigenvalue error**:

$$[C_1, \ldots, C_m] = \min_{[C_1,\ldots,C_m]} \max_{j=1,\ldots,m} |\lambda_j - \hat{\lambda}_j|$$

where C_i are the modules, λ_j are the eigenvalues of the transition matrix of the original process and $\hat{\lambda}$ are the eignvalues of the transition matrix of the milestoning process.
The minimization is done using simulated annealing, the initial input is obtained by using the PCCA+ algorithm.

Relevant papers: *Djurdjevac, N., Bruckner, S., Conrad, T., Schuette, Ch.. Random walks on complex modular networks. submitted.*
Djurdjevac, N., Sarich, M., Schuette, Ch. (2010) Proceedings of ICM 2010

Detected modules



After identifying the modules, we turn to study the communication between them. We are interested in the paths and nodes through which the modules communicate.To this end, we look at all the possible different **settings** of communication between every single module and the union of all other modules.

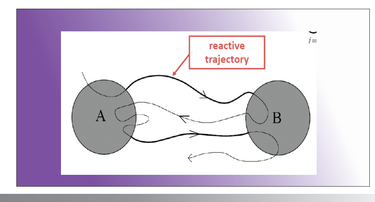
(4) Calculating statistical properties.

Key nodes are nodes through which much of the random walker "traffic" goes through as it traverses between the modules.
Those nodes take part in the most intensive communication in the network. To identify these nodes we use **TPT**. We study the **reactive trajectories** between each module A and the union of the rest of the modules B: those are the "pieces" of the random walk that begin at A and end at B.
For this we use the committors, defined as:

$$q(x) = P[\tau_x(A) < \tau_x(B)]$$

for some subsets A, B of the network, where τ is the hitting time.
The committor gives the probability that the walker, starting in x, enters A earlier than B.

Relevant paper: *Metzner, P., Schuette, Ch., Vanden-Eijnden, E. Transition path theory for Markov jump processes. Multiscale Modeling and Simulation, 7(3):1192–1219, 2009.*



(5) Computing effective current

Using the committors we can now define the **probability current:** the rate f_{ij}^{AB} at which reactive trajectories flow from node i to node j. where μ is the invariant measure.

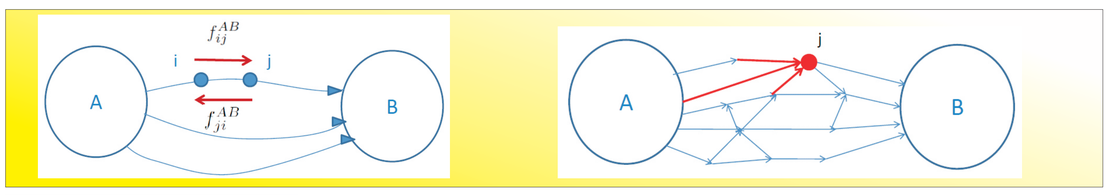
$$f_{ij}^{AB} = \begin{cases} \mu(i)\, q(i)\, p(i,j)\, (1 - q(j)), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

Every edge of the network is now parameterized by its **effective current**

$$f_{ij}^{+} = \max\left(f_{ij}^{AB} - f_{ji}^{AB}, 0\right)$$

For a single node j we look at the reactive flow through j:

$$k_j = \sum_{i \in P_j} f_{ij}^{+} = \sum_{i \in S_j} f_{ji}^{+} \qquad P_j = \{i \in V : \quad f_{ij}^{+} > 0\}, \quad S_j = \{i \in V : \quad f_{ji}^{+} > 0\}$$



(6) Scoring the key nodes.

We describe the **global transition** behavior between sets A and B. This is the average number of reactive trajectories, or the average number of transitions from A to B per time unit.

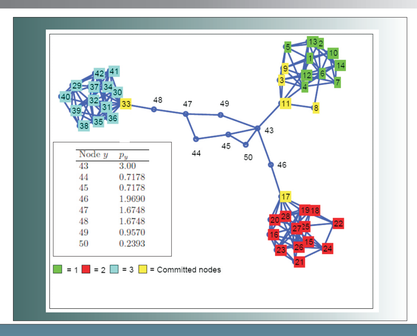
$$k_{AB} = \sum_{i \in A, j \in V} f_{ij}^{+}$$

To identify the key nodes, we now compute for each candidate node j the **rate of reactive trajectories** that go through j when $A \to B$.

$$p_j^{AB} = \frac{k_j}{k_{AB}}$$

The figure to the right shows the key nodes of our sample network, with the 3 different settings, and a table giving the scores for each.

Relevant paper: *Djurdjevac, N., Bruckner, S., Conrad, T., Schuette, Ch.. Random walks on complex modular networks, submitted.*

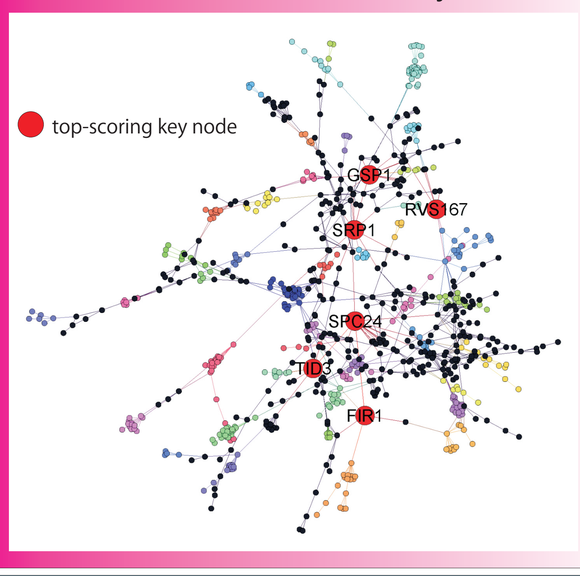


Results

We apply our methods to the filtered yeast network [1] (FYI) , where the authors constructed a high-confidence PPI network and analyzed its hubs.
Our analysis gave 43 modules, visualized on the figure on the right. Testing the functional enrichment of our modules provided good results, as our modules comprise, for example, the Arp2/3 protein complex, the 20S proteasome, all proteins involved in the processs of double-stranded DNA binding, and those annotated with RNA polymerase III transcription factor activity, along with other significant annotations.
We also looked at key nodes, graphing the high-scoring ones on the right. Our key nodes show high correspondence with the essential hubs identified in [1].

We plan to continue our research into the biological nature of the hubs, and also to study the effects of the various parameters and the identification of the spectral gap on our results. We are extending our results to the study of other networks, such as social and transportation networks.

Detected modules and key nodes



[1] Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004;430:88–93.