Heat Flow-Thermodynamic Depth Complexity in Networks

Francisco Escolano, Miguel A. Lozano
University of Alicante
{sco, malozano}@dccia.ua.es

Edwin R. Hancock
University of York
erh@cs.york.ac.uk

Abstract

In this paper we establish a formal link between network complexity in terms of Birkhoff-von Neumann decompositions and heat flow complexity (in terms of quantifying the heat flowing through the network at a given inverse temperature). Furthermore, we also define heat flow complexity in terms of thermodynamic depth, which results in a novel approach for characterizing networks and quantify their complexity. In our experiments we characterize several protein-protein interaction (PPI) networks and then highlight their evolutionary differences.

1. Introduction

The quantification of the intrinsic complexity of networks has attracted significant attention, in a number of fields including complexity science, pattern recognition and machine learning, due to its fundamental practical importance. Some complexity characterizations rely on spectral graph theory (see [2] for applications in computational biology, [7] for biological, social and other kinds of networks, and [6][5] for applications to pattern recognition). The recent extensive use of spectral theory is due to: a) it explains some previous approaches (number of spanning trees, path-length distribution, clustering, and so on) from a different perspective and b) it is flexible enough to allowing the development of new characterizations. In this paper we explore the connection between convex polytopes (and those of the Birkhoff type in particular), heat kernels in graphs, the well known thermodynamic depth approach to complexity, and network complexity itself. Some work in this direction has been done recently [6], but no formal connections between polytopes and heat-flow characterization of structural entropy [6] has been developed so far. Our main contribution here is to formally specify the complexity profiles of both approaches to structural complexity, showing that they have a qualitative simi-
lar behavior and that the complexity corresponding to the maximum entropy (ME) Birkhoff-von Newman decomposition is derived from that corresponding to the maximum flow. Thus, the phase-transition point always exists and it is characterized by such maxima. Moreover we establish links between heat flow complexity and thermodynamic depth. We also apply this characterization to PPI networks.

2. Polytopal vs Heat Flow Complexity

Theorem 1 (Birkhoff-von Neumann (BvN) [3]). Let \( B_n \) be the set of doubly stochastic matrices \( B = [b_{ij}]_{n \times n} \) of dimension \( n \times n \) (Birkhoff polytope). Then every doubly stochastic matrix (DSM) \( B \) can be expressed as a convex combination of permutation matrices (PM):

\[
B = \sum_{\alpha} p_{\alpha} P_{\alpha}, \ \forall B \in B_n \text{ and } \sum_{\alpha} p_{\alpha} = 1 \quad \text{s.t.} \quad \sum_{\alpha} p_{\alpha} \geq 0 \quad \forall \alpha.
\]

Thus \( B_n \) is the convex hull of the set of the \( n \times n \) permutation matrices. However, the representation of a DSM in terms of many PMs is not unique because \( B_n \) is not a simplex. The barycenter of \( B_n \) is the van der Waerden constant matrix \( B_\alpha \) with all entries equal to \( 1/n \).

Theorem 2 (Agrawal, Wang & Ye [1]). The maximum entropy (ME) BvN decomposition of a DSM \( B \) is the solution to the left optimization problem below (primal) whose dual one is on the right:

\[
\min_{S_n} \sum_{\alpha} p_{\alpha} (\log p_{\alpha} - 1) \quad \text{max } \quad B : Y - \sum_{\alpha} e^{Y : P_{\alpha}}
\]

\[
s.t. \quad \sum_{\alpha} p_{\alpha} P_{\alpha} \leq B \quad s.t. \quad \sum_{\alpha} e^{Y : P_{\alpha}} P_{\alpha} \leq B
\]

\[
p_{\alpha} \geq 0 \quad 0 \geq Y_{ij} \geq -n \log n / b_{\min} \quad \forall i, j
\]

where \( S_n \) is the set of permutations of \( \{1, 2, \ldots, n\} \), \( X : Z = \sum_{ij} X_{ij} Z_{ij} = \text{trace}(XZ^T) \) is the Frobenius inner product, \( Y \in R^{n \times n} \), a matrix of Lagrange multipliers each corresponding to one constraint (component) in \( B = \sum_{\alpha} p_{\alpha} P_{\alpha} \) and \( b_{\min} = \min\{B_{ij}\} \).
In [1] it is shown how to solve approximately the dual of the ME problem. In practice, however, instead of finding a unique representation for $B$ it is preferable to obtain greedily just one of them. To that end, the constructive proof of the BvN theorem is used. This is the origin of polytopal complexity [6].

**Definition 1** (Polytopal Complexity [6]). Given $G = (V,E)$, an undirected and unweighted graph with diffusion kernel $K_\beta(G)$, and BvN decomposition $K_\beta(G) = \sum_{\alpha=1}^n p_\alpha P_\alpha$, we define the polytopal complexity of $G$ as the $\beta$-dependent function

$$B\mathcal{C}_\beta(G) = \frac{H(\mathcal{P})}{\log_2 n} = \frac{\log_2 \gamma + D(\mathcal{P}||\mathcal{U}_n)}{\log_2 n}, \quad (1)$$

where $\mathcal{P} = \{p_1, \ldots, p_n\}$ is the probability density function (pdf) induced by the decomposition, $H(.)$ the entropy and $D(.)$ the Kullback-Leibler divergence $D(\mathcal{P}||\mathcal{Q}) = \sum_\alpha p_\alpha \log \frac{p_\alpha}{q_\alpha}$.

In [6] it is argued that the typical signature is heavy tailed, monotonically increasing from 0 to $\beta^+ \equiv \arg \max \{B\mathcal{C}_\beta(G)\}$ and either monotonically decreasing or stable from $\beta^+$ to $\infty$ where $B\mathcal{C}_\beta(G) = 1$ is reached. Thus, $\beta^+$ represents the most significant topological phase transition regarding the impact of the diffusion process in the topology of the input graph. However, no characterization theorem has been enounced so far in order to validate the latter assumptions. In addition, in [5] it is showed that the $O(n^3)$ computational complexity of the greedy BvN decomposition for each $\beta$ precludes the use of the descriptor for the practical analysis of complex networks. Thus, a new descriptor, qualitatively similar but more efficient than the current one, and also providing a simpler analytical framework, is needed.

**Definition 2** (Heat Flow Complexity [5]). Given $G = (V,E)$ with $|V| = n$ and adjacency matrix $A$. The diffusion kernel is $K_\beta(G) = \exp(-\beta\mathcal{L}) \equiv \Phi \Lambda \Phi^T$, being $\Lambda = \text{diag}(e^{-\beta \lambda_1}, e^{-\beta \lambda_2}, \ldots, e^{-\beta \lambda_n})$, and $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ are the eigenvalues of $\mathcal{L}$. Therefore, the total heat flowing through the graph at a given $\beta$ is:

$$F_{\beta}(G) = \sum_{i=1}^n \sum_{j \neq i} A_{ij} \left( \sum_{k=1}^n \phi_k(i) \phi_k(j) e^{-\beta \lambda_k} \right), \quad (2)$$

where $A_{ij} = 1$ iff $(i,j) \in E$. Then, heat flow complexity is defined as:

$$F\mathcal{C}_\beta(G) = \frac{\log_2 (1 + F_{\beta}(G))}{\log_2 n}, \quad (3)$$

3. Characterization of Polytopal and Flow Complexity

**Definition 3** (Phase-Transition Point) Let $G = (V,E)$ with $|V| = n$, edge-set $E$ and a finite inverse temperature $\beta^+ \geq 0$. A diffusion kernel $K_{\beta^+}(G)$ on graph $G$ has a phase transition point or PTP at $\beta^+$ which is the maximal inverse temperature for which the sum of the off-diagonal elements of the diffusion kernel (or Gram matrix) is less that the sum of the on-diagonal elements. In other words, $\sum_{i=1}^n \sum_{j \neq i} K_{\beta^+}(i,j) < \text{trace}(K_{\beta^+})$ and also $\forall \beta > \beta^+: \sum_{i=1}^n \sum_{j \neq i} K_{\beta^+}(i,j) \geq \text{trace}(K_{\beta^+})$.

It is straightforward to prove that the PTP always exists and it is unique. The analysis of the limiting cases: $K_0 = I_n$ and $K_\infty = \mathcal{B}$, where $\Xi_\beta = \text{trace}(K_{\beta}) - \sum_{i=1}^n \sum_{j \neq i} K_{\beta}(i,j)$ yields $\Xi_0 = n$ and $\Xi_\infty = -n$. Actually $-n$ may be reached as soon as the kernel converges to $\mathcal{B}$ (reaches the equilibrium point). Local maxima of $\Xi_\beta$ are precluded by the monotonic nature of the diffusion process and therefore $\Xi_\beta$ is a monotonically decreasing function with a minimum at equilibrium. The existence of a unique PTP is key to relating heat flow and maximal entropy.

**Theorem 3** (Phase-Transition) Let $\beta^+ > 0$ define a PTP. Then, the heat flow $F_{\beta^+}(G)$ is maximal among all choices of $\beta$. This implies that when $K_{\beta^+}(G) = \sum_{\alpha=1}^n p_\alpha P_\alpha$ and the maximum entropy BvN decomposition yields $H_{\beta^+}(\mathcal{P})$ with $\mathcal{P} = \{p_1, \ldots, p_n\}$, then $H_{\beta^+}(G)$ is maximal over $\beta$.

The proof of the latter theorem, exploits Th. 2, but its detailed description is beyond the scope of this contribution due to space constraints (see the corresponding author website for more details). Anyway, the key fact is that Th. 3 establishes a correspondence between PTPs and maximum entropy BvN decompositions.

4. Heat Flow - Thermodynamic Depth Complexity

The application of thermodynamic depth (TD) to characterize network complexity demands the formal specification of the micro-states whose history leads to the macro-state (of the network). Here we define such micro-states in terms of expansion subgraphs.

**Definition 4** (Node History & Expansion Subgraphs) Let $G = (V,E)$ with $|V| = n$. Then the history of a node $i \in V$ is $h_i(G) = \{e(i), e^2(i), \ldots, e^n(i)\}$ where:

$e(i) \subseteq G$ is the first-order expansion subgraph given by

\[1\text{http://sites.google.com/site/scohomepage/}]}
i and all j ∼ i, e^2(i) = e(e(i)) ⊆ G is the second-order expansion consisting on z ∼ j : j ∈ V_e(i), z /∈ V_e(i), and so on until p cannot be increased. If G is connected e^p(i) = G, otherwise e^p(i) is the connected component to which i belongs.

Every h_i(G) defines a different causal trajectory leading to G, if it is connected, or to one of its connected components otherwise. Thus, in terms of TD the full graph G or the union of its connected components is the macro-state (macroscopic state). The depth of such a macro-state relies on the variability of the causal trajectories leading to it. The higher the variability, the more complex is to explain how the macro-state is reached and the deeper is this state. Therefore, in order to characterize each trajectory we combine the heat flow complexities of its expansion subgraphs by means of defining minimal enclosing Bregman balls (MEBB) [9]. Bregman divergences D_F define an asymmetric family of similarity measures, each one characterized by a strictly convex generator function F : X → R^+, where X ⊆ R^d is a convex domain, and d the data dimension (in this case the number of discretized β - inverse temperatures). Given two patterns (discretized functions in this case) f and g, D_F(f[|g]) = F(f) - F(g) - (f - g)^T ∇ F(f). Here, we use the I-Kullback-Leibler divergence D_F(f[|g]) = \sum_{i=1}^d f_i log \frac{f_i}{g_i} - \sum_{i=1}^d f_i + \frac{1}{\sum_{i=1}^d g_i} with F(f) = \sum_{i=1}^d (f_i log f_i - f_i) (un-normalized Shannon entropy) which yields better results (more representative centroids of heat flow complexities) than other divergences/distances like that of Itakura-Saito one. When using the I-KL divergence in R^d, we have that ∇ F(f_i) = log f_i and also that ∇^{-1} F(f_i) = e^{f_i} (obviously natural logarithm is assumed). Using these formal ingredients we define the causal trajectory in terms of MEBBs.

**Definition 5** (Causal Trajectory) Given h_i(G), the heat flow complexity f_i = f(e^i(i)) for the t – th expansion of i, a generator F and a Bregman divergence D_F, the causal trajectory leading to G (or one of its connected components) from i is characterized by the center \bar{c}_i ∈ R^d and radius r_i ∈ R of the MEBB B^{\bar{c}_i,r_i} = \{ f_i ∈ X : D_F(\bar{c}_i || f_i) ≤ r_i \}.

Solving for the center and radius implies finding \bar{c}_i, r^*_i minimizing r s.t. D_F(\bar{c}_i || f_i) ≤ r ∀ f_i ∈ X with |X| = T. Considering the Lagrange multipliers α_i we have that \bar{c}_i = ∇^{-1} F(\sum_{t=1}^T α_i f_i ∇ F(\bar{f}_i)). The efficient algorithm in [9] estimates both the center and multipliers. This is closely related to Core Vector Machines [10], and it is quite interesting to focus on the non-zero multipliers (and their support vectors) used to compute the optimal radius. More precisely, the multipliers define a convex combination and we have α_i ∝ D_F(e^α || \bar{f}_i), and the radius is simply chosen as: r^* = \max_{α_i > 0} D_F(e^α || \bar{f}_i).

**Definition 6** (TD Network Depth) Given G = (V, E), with |V| = n and all the n pairs (c_i, r_i), the heat flow thermodynamic depth complexity of G is characterized by the MEBB B^{c,r} = \{ c_i ∈ X_i : D_F(c_i || c_i) ≤ r \} and D_{min} = \min_{f ∈ B^{c,r}} D_F(f^∞||f), where f^∞ = f(B_v) ∈ R^d is the van der Waerden complexity trace. Then, the TD depth of network is given by D(G) = r × D_{min}.

The latter definitions of complexity and depth are highly consistent with summarizing node histories to find a global causal trajectory the more tightly bounded as possible. Here, r quantifies the historical uncertainty: the smaller r the simpler (shallower) is G. However, this is not enough for structures because many networks with quite different complexities may have the same r. Therefore, we define the depth of the network complementing randomness as suggested in the thermodynamic depth approach. In our case, the projection of f^∞ on the MEBB preserves the definition of entropy in terms of the distance to the uniform distribution. The combinations or hierarchies of MEBBs have proved to be more effective than ball trees for nearest-neighbor retrieval [4]. In the computation of depths, the Legendre duality (convex conjugate) is key because it establishes a one-to-one correspondence between the gradients ∇ F and ∇ F^T due to the convexity of F. Therefore, the Bregman projection f of f^∞ on the the border of B^{c,r} lies on the curve f^0 = θ∇ F(c) + (1 − θ)∇ F(f^∞) with θ ∈ [0, 1] and f^0 = ∇^{-1} F(f^0). The projection f be easily found (approximately) through bisection search on θ.

5. TD of PPIs and Conclusions

The experimental section relies on PPIs extracted from STRING–8.22 (see Fig. 1-left). In a first experiment, we consider PPIs related to histidine kinase, a key protein in the development of signal transduction, corresponding to 10 species belonging to 10 phyla of Bacteria. We select subjectively 3 PPIs (simple, complex and more-complex) from each species and compute their TDs. In 70% of the cases, TD matches intuition. When comparing with Estrada’s spectral homogeneity descriptor [7] we also find that the ratio between intraclass and interclass variability is slightly better (smaller) for TD (0.6840 vs. 0.7383). The
second experiment consists of analyzing 224 PPIs, also related to histidine kinase, from 6 different groups (all the PPIs in the same group corresponds to the same species) with the following evolutive order (from older to more recent): Aquifex—4 PPIs, Thermotoga—4 PPIs, Gram-Positive—52 PPIs, Cyanobacteria—73 PPIs Proteobacteria—45 PPIs. There is an additional class (Acidobacteria—46 PPIs). Histogramming TDs reveals typically long tailed distributions with most of the TDs concentrated at a given point. Are these points ordered according to the evolutive order? This question can be answered by studying the cumulative distributions instead of the pdfs (Fig. 1-right/top). In such case, reaching the top (cumulative=1) soon indicates low TD whereas reaching it later indicates high TD. Then, it can be seen that the evolving complexity of the signal transduction mechanism driven by the histidine kinase is properly quantified by TD for the 5 first phyla studied. However, the Acidobacterium sp. chosen seems older than Gram-Positive which seems not to be the case 3. In the bottom of Fig. 1-right we show some $\vec{c}_i$s of all classes, and their intraclass variability is low (similar shape). Thus, we can conclude that TD is a good principled tool for analysing the complexity of networks, and PPIs in particular, and multi-species experiments are planned for a near future.

Acknowledgements

F. Escolano and M.A. Lozano are supported by project TIN2008-04416 of the Spanish Government. E. Hancock is supported by a Royal Society Wolfson Research Merit Award and the EU FET project SIMBAD (213250).

References


3In this study we are only considering a species per class.