Finding role communities in directed networks using Role-Based Similarity, Markov Stability and the Relaxed Minimum Spanning Tree

Mariano Beguerisse-Díaz
Department of Mathematics
Imperial College London
London, SW7 2AZ
United Kingdom
m.beguerisse@imperial.ac.uk

Borislav Vangelov
Department of Mathematics
Imperial College London
London, SW7 2AZ
United Kingdom
borislav.vangelov09@imperial.ac.uk

Mauricio Barahona
Department of Mathematics
Imperial College London
London, SW7 2AZ
United Kingdom
m.barahona@imperial.ac.uk

***Abstract*—We present a framework to cluster nodes in directed networks according to their roles by combining Role-Based Similarity (RBS) and Markov Stability, two techniques based on flows. First we compute the RBS matrix, which contains the pairwise similarities between nodes according to the scaled number of in- and out-directed paths of different lengths. The weighted RBS similarity matrix is then transformed into an undirected similarity network using the Relaxed Minimum-Spanning Tree (RMST) algorithm, which uses the geometric structure of the RBS matrix to unblur the network, such that edges between nodes with high, direct RBS are preserved. Finally, we partition the RMST similarity network into role-communities of nodes at all scales using Markov Stability to find a robust set of roles in the network. We showcase our framework through a biological and a man-made network.**

I. Introduction

Among the many systems that can be formalised as networks, there are important examples where the directionality of the network is crucial, e.g., the web, ecological systems, information and transport networks. However, directionality brings in subtle mathematical complexities and is often neglected in many approaches for network analysis. Such directed networks naturally lend themselves to be analysed from the perspective of *flows*. An important aspect of directed networks is the notion of *roles*, e.g., leader vs follower or hub vs authority. Here we show that a nuanced classification of nodes in terms of their role in the network may be obtained from the analysis of directed flows. In other words, we seek to find nodes that are similarly positioned in the network–with respect to flows–and obtain broad categories into which they can be classified. In this paper, we present a method to find role clusters in directed networks based on the analysis of flow patterns, and show examples of its application to a selection of networks. Section II contains a brief introduction and references to the specific techniques we use and an overview of the method. Section III provides examples of our method.

II. Methodology

Let $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ be an unweighted and directed network with node set $\mathcal{N}$, $|\mathcal{N}| = N$, edge set $\mathcal{E}$, and adjacency matrix $\mathbf{A}$ where $a_{i,j} = 1$ denotes the existence of a directed edge from node i to j. Each node has in-degree k_{in} (the number of nodes that link to it) and out-degree k_{out} (the number of nodes to which it links). The $N \times 1$ vectors of in- and out-degrees are denoted by $\mathbf{k}_{in}$ and $\mathbf{k}_{out}$.

We find the role-communities of $\mathcal{G}$ following three steps:

1) From the adjacency matrix A, construct a $N \times N$ node similarity matrix based on the directed connectivity profile of the nodes, as given by RBS (Sec. II-A).
2) From this RBS matrix, obtain a (new) undirected similarity network using the RMST algorithm such that two nodes are connected if their connectivity profiles are highly similar (Sec. II-B).
3) Find robust partitions of this RMST similarity network into communities of nodes with the same roles at several levels of resolution using Markov Stability, a multiscale community detection method (Sec. II-C).

A. Role-Based Similarity in directed networks

In a directed network the in- and out-connectivities of the nodes contain information about the role of each node in the network. The simplest categorisation of nodes into "leaders" and "followers", according to the predominance of their in- or out-degree, is often illustrative but limited as it neglects the full topology and complexity of the network. Other methods that harness further information from the network structure can be used to compute the "status" index [1], PageRank [2], or the "Hub"/"Authority" score [3]. Though powerful, these methods are limited by the fact that they split the nodes into at most two categories (or further categories according to a one-dimensional classification).

To go beyond the 'leader-follower' dichotomy, we employ Role-Based Similarity (RBS) [4], [5], a method that calculates how similar nodes are to each other in terms of the scaled number of adjacent directed paths of *all meaningful lengths* (i.e., no longer than N). The idea is to create a $1 \times 2K_{max}$ feature vector for each node, $\mathbf{x}_i$, whose entries are the weighted number of paths of lengths from 1 to $K_{max} < N$ originating

and ending in node i. All the feature vectors $\mathbf{x}_i$ are stored as the rows of the $N \times 2K_{max}$ matrix:

$$\mathbf{X} = \left[\ldots, \left(\beta\mathbf{A}^T\right)^k \mathbb{1}, \ldots \middle| \ldots, \left(\beta\mathbf{A}\right)^k \mathbb{1}, \ldots\right], \qquad (1)$$

where $k = 1, \ldots, K_{max}$. Note that the number of originating paths of length k from all nodes is given by $\left(\beta\mathbf{A}\right)^k \mathbb{1}$, and the number of arriving paths $\left(\beta\mathbf{A}^T\right)^k \mathbb{1}$, where $\mathbb{1}$ is the $N \times 1$ vector of ones. Here, we use $\beta = \alpha/\lambda_1$, where λ_1 is the largest eigenvalue of $\mathbf{A}$ and $\alpha \in (0, 1)$, which assures convergence of the sequence $\beta^k \mathbf{A}^k$ as $k \to \infty$. Hence the columns of $\mathbf{X}$ contain the number of in- (or out-) paths of length k for each node weighted by β^k.

In addition to guaranteeing convergence, the parameter α also determines the weight given to each path length: smaller values of α give more weight to shorter paths than to longer ones. If $\alpha \ll 1$, the columns of $\mathbf{X}$ converge rapidly (because $\lim_{k\to\infty} \left(\beta\mathbf{A}\right)^k \mathbb{1} = \mathbf{0}$), which results in feature vectors based only on local properties based on short paths (i.e., in the limit $\alpha \to 0$, the feature vector only contains k_{in} and k_{out}). As α is increased, we incorporate more global features of the network in our analysis. Results in [4] indicate that $\alpha = 0.95$ provides a good balance between the information gathered from the local and global flow structure in the network. However, the systematic determination of α for each network is currently the focus of further investigation.

From $\mathbf{X}$ we then compute the RBS matrix $\mathbf{Y}$, whose entries contain the cosine-similarity between all rows of $\mathbf{X}$:

$$y_{i,j} = \frac{\mathbf{x}_i \mathbf{x}_j^T}{||\mathbf{x}_i||_2 \, ||\mathbf{x}_j||_2}. \qquad (2)$$

When nodes i and j have an *identical* pattern of path flows *at all lengths* in $\mathcal{G}$, then $\mathbf{x}_i$ and $\mathbf{x}_j$ are collinear and $y_{i,j} \simeq 1$. On the contrary, when nodes i and j do not have any number of paths in common at any length (e.g., when i is a source and j a sink node) then $y_{i,j} = 0$. The RBS matrix $\mathbf{Y}$ is symmetric, usually full, and could be used to find groups of nodes with similar connectivity. However, as is usually the case with correlation or distance matrices, clustering $\mathbf{Y}$ directly is problematic because of its lack of sparsity and the unstructured nature of geometric distances in high-dimensional spaces. To unblur the structure of $\mathbf{Y}$, we extract a similarity network that select links between nodes with strong similarity while discarding weak similarities that can be explained in terms of other relationships in the network, as we explain now.

B. Obtaining the similarity network from the RBS matrix

The N feature vectors containing the flow profiles of the nodes are defined in a high-dimensional space of $2K_{max}$ dimensions. However, because the coordinates of the vectors are smoothly related to each other, we expect that the vectors of all nodes will lie in a lower dimensional manifold whose structrure can be well captured by a graph with a geometric structure. Here we use the Relaxed Minimum-Spanning Tree (RMST) algorithm, a method that incorporates local and global features of the data to recover such a network from the RBS matrix.

First, we define the 'dissimilarity' (or 'distance') matrix $\mathbf{Z}$, with $z_{i,j} = 1 - y_{i,j}$, i.e., the more similar i and j are to each other, the smaller the value of $z_{i,j}$ and the closer i and j lie. The RMST algorithm constructs a network with adjacency matrix $\mathbf{E}$ from $\mathbf{Z}$ as follows. First, consider $\mathbf{Z}$ to be the adjacency matrix of a weigthed graph and obtain a Minimum Spanning Tree (MST) in it, setting $e_{i,j} = 1$ if nodes i and j are neighbours in the tree. Each node pair (i, j) is connected by a path (or sequence of edges) $\{(i, k), (k, h), \ldots, (m, j)\}$ in the MST. We then find the maximal weight in $\mathbf{Z}$ along the MST path:

$$\text{mlink}_{ij} = \max\{z_{i,k}, z_{k,h}, \ldots, z_{m,j}\}.$$

If mlink_{ij} is significantly smaller than $z_{i,j}$ then the MST-path is considered to be a good model to explain the similarity between nodes i and j and discard the direct link between them, i.e., we leave $e_{i,j} = 0$. If $z_{i,j}$ is comparable to mlink_{ij} then there is not sufficient evidence to believe that the MST-path is a better model and we include the direct link $e_{i,j} = 1$. More precisely, we set $e_{i,j} = 1$ when

$$\text{mlink}_{ij} + \gamma(d_i + d_j) > z_{i,j}, \qquad (3)$$

where $d_i = \min_k z_{i,k}$ and γ is a parameter ($\gamma = 0.5$ here). The term γd_i approximates the local distribution of points (in $\mathbf{Z}$) around i and is motivated by the Perturbed Minimum Spanning Tree algorithm [6].

The RMST similarity network is an unweighted, undirected graph where two nodes are connected only if their flow feature vectors are highly similar, *regardless of whether they are neighbours in the original graph $\mathcal{G}$ or not*. We can also obtain a weighted similarity graph by Hadamard-multiplying $\mathbf{E}$ and $\mathbf{Y}$. The RMST network is sparse if the data in $\mathbf{Y}$ results from a local geometric structure (which the RMST tries to recover), and is more amenable to analysis using network analysis techniques such as the community detection method we describe below.

C. Role-communities with Markov Stability

Community detection in networks has been studied extensively and there exist a wide variety of methods, each with their own advantages [8]. Here, we use the method known as Markov Stability [9], [10] to detect 'role-communities' in the RMST similarity network. There are a number of advantages to using Markov Stability in this case, key among them is the ability to detect communities in the network at *all scales* via a continous-time Markov process evolving in time. Due to this dynamic zooming, Markov Stability does not impose an *a priori* number of roles (i.e., role-communities) in the network, but rather detects the presence of robust partitions at all levels of resolution. Hence we learn the number of roles by exploring the network with a continuous-time diffusion process and finding robust, optimised partitions across scales.

Consider $\mathbf{E}$, the adjacency matrix of the undirected RMST similarity network, which by construction is connected. Define $\mathbf{k} = \mathbf{E}\mathbb{1}$, the vector of degrees, and $\mathbf{D} = \text{diag}(\mathbf{k})$ so that $\mathbf{D}^{-1}\mathbf{E}$ is a row-stochastic matrix. The normalised Laplacian

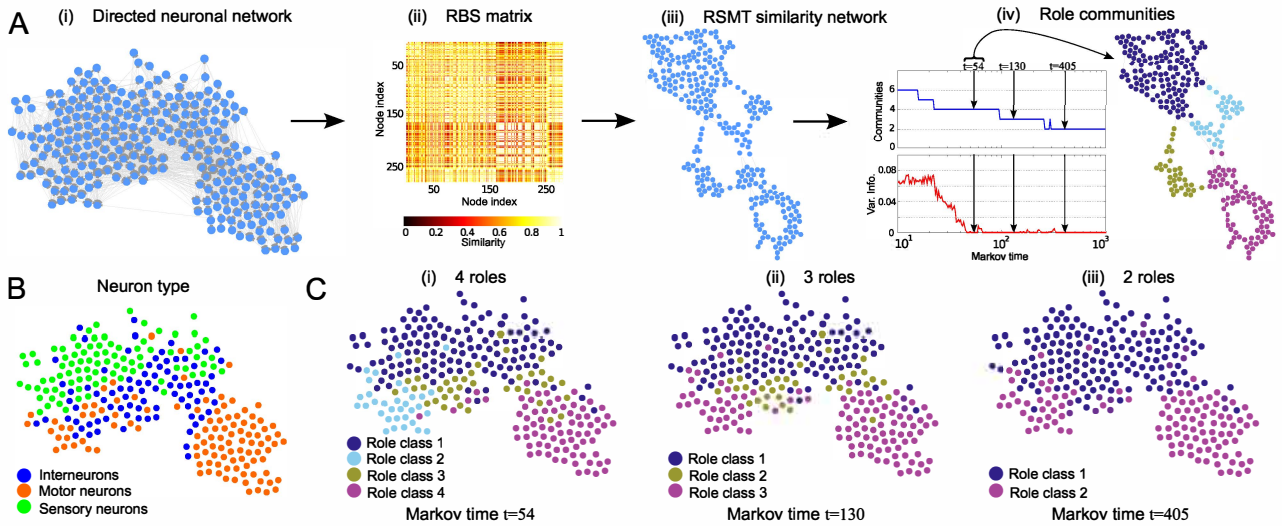

Fig. 1. Detecting role communities in the *C. elegans* neural network. **A**: *(i)*: Directed neural network [7]. *(ii)*: Heatmap of the RBS matrix ($\mathbf{Y}$, Sec. II-A): the higher the similarity between two nodes, the lighter the cell. *(iii)*: Similarity network obtained from the RBS matrix with the RMST algorithm (Sec. II-B): only nodes with highly similar in- and out-flow patterns are connected. *(iv)*: Community detection of the RMST similarity network using Markov Stability (Sec. II-C): robust partitions into 4, 3 and 2 role-communities are detected at different levels of resolution. **B**: Types of neurons in the *C. elegans* network, as given in Ref. [7]. **C**: The role-classes obtained using RBS+RSMT+Stability at different levels of resolution (Markov times).

of the system is then $\mathbf{L} = \mathbf{I}_N - \mathbf{D}^{-1}\mathbf{E}$, and the transition matrix of the continuous-time Markov process of duration $t > 0$ (the Markov time) is $P(t) = e^{-t\mathbf{L}}$ (giving the probability of transitioning form node i to j in a process of duration t) [11]. A hard partition of the network into C groups of nodes can be encoded in a $N \times C$ matrix $\mathbf{H}$ (i.e., $\mathbf{H}_{i,c} = 1$ is node i belongs to community c, and $\sum_c \mathbf{H}_{i,c} = 1 \, \forall \, i$). The *Markov Stability* of the partition at time t is defined as the trace of the clustered autocovariance of the diffusion process [9]:

$$r(t, \mathbf{H}) = \text{trace}\left(\mathbf{H}^T \left[\Pi P(t) - \pi\pi^T\right] \mathbf{H}\right), \qquad (4)$$

where π is the steady-state distribution of the process, and $\Pi = \text{diag}(\pi)$. We find the communities in the similarity network for a given t by maximising $r(t, \mathbf{H})$ over the space of partitions; that is, we find the network partitions that maximise the retention of flows over a timescale.

Maximising equation (4) is an NP-hard problem, with no guarantees of global optimality. The optimised partitions are found using the Louvain algorithm [12], a greedy heuristic that has been shown to give good results in practice. To find the relevant partitions of the network at any time scale, we use a robustness criterion based on the consistency of the optimisation quantified through an information-theoretical measure. At each Markov time, we obtain optimised partitions of the network by running the Louvain method 100 times, each time using a random initial guess. To gauge the robustness of the set of optimised partitions, we calculate the mean pairwise Variation of Information (VI) of the ensemble of Louvain solutions [10]. The VI between two partitions $\mathbf{H}_1$ and $\mathbf{H}_2$ is [13]:

$$VI(\mathbf{H}_1, \mathbf{H}_2) = \frac{1}{\log N}\left(2H(\mathbf{H}_1, \mathbf{H}_2) - H(\mathbf{H}_1) - H(\mathbf{H}_2)\right),$$

with $H(\mathbf{H}) = -\sum_c p(c) \log p(c)$ and $p(c) = \sum_i \mathbf{H}(i,c)/N$. When the optimisation algorithm finds partitions of the network that are consistently similar (a hallmark of robust community structure), the mean pairwise VI is low; when there is no clear community structure the optimisation produces partitions that are different to each other, resulting in a high mean VI. Finally, to make sure we detect all the relevant role-communities, we optimise equation (4) for all Markov times, keeping track of the mean VI as a function of t, and detecting communities that are also persistent across Markov times. We now provide examples of finding node roles in different networks using the RBS/RMST/Markov Stability methodology explained in this section.

III. Examples

A. C. elegans neural network

The directed neural network of *C. elegans* records the chemical synapses and the junctions between 279 neurons [7]. Figure 1A shows the steps of the analysis: the original *unweighted* directed neuronal network; computation of the RBS matrix $\mathbf{Y}$ with $\alpha = 0.95$ and $K_{max} = 116$; generate the RMST similarity network; and find role-communities in it using Markov Stability. As shown in Fig. 1A-(iv), our analysis finds meaningful partitions of the RMST network into up to four role-classes.

C. elegans is known to have three types of neurons: sensory, motor, and interneurons [7], shown in Figure 1B with different colours. We display the neural network on the plane as in Ref. [7]: the horizontal axis corresponds to the entries of the Fiedler vector reflecting mostly body position, and the vertical axis corresponds to processing depth with respect to information flow. Fig. 1C shows that the partition into four, three and two roles broadly reflects the biological groups.

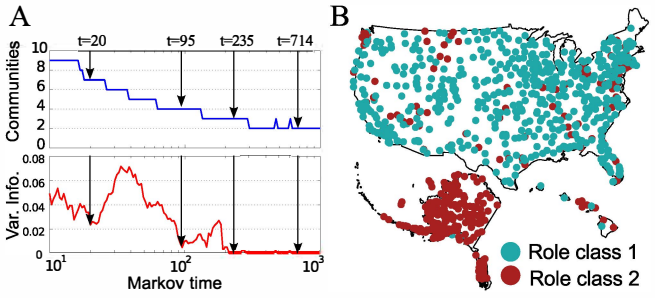

Fig. 2. Roles in the US airport network. **A**: Number of role-communities found in the RMST similarity network at all Markov times (top) and the Variation of Information of the partitions (bottom). **B**: The two roles found in the US airport network at Markov time $t = 714$.

Among the four roles, we find one group (Role 1, dark blue) which corresponds mostly to sensory neurons and some interneurons. Role 3 is formed by a subset of the interneurons. Interestingly, motor neurons are split in two classes (Roles 2 and 4) clearly separated along the body of the worm (x-axis). The motor neurons in Roles 2 and 4 are merged into a common role when we cluster the RMST network into 3 role-communities. Finally, the partition into two roles separates the groups along the lines of sensory and motor neurons—interneurons are split in both, with slightly more interneurons in the sensory group.

B. US airport network

We also investigated the roles in the unweighted network of $N = 957$ airports in the United States [14], [15]. The analysis proceeds as before and we calculate the RBS matrix with $\alpha = 0.92$ and $K_{max} = 78$. Figure 2A shows the number of roles found at different levels of resolution. We find partitions into seven or fewer role classes—the VI has pronounced dips at $t = 20$, $t = 95$, $t = 235$, and $t = 714$, corresponding to 7, 4, 3 and 2 role-communities. Interestingly, there is always a distinctive role for a large group of Alaskan airports across all levels of resolution which persists separately up to the highest level of resolution. As shown in Fig. 2B, where we show the US map with nodes coloured according to the two role classes at $t = 714$, the most striking attribute is that practically all airports in Alaska (including the two largest, Anchorage and Fairbanks) belong to role class 2. Transport in Alaska, a large and sparsely populated region with many remote settlements scantily connected by roads, relies on local airports and airstrips. These ingredients contribute to create a distinct (and less reciprocal) air-transportation connectivity, which sets Alaska apart from most of the rest of the US. The few nodes in the mainland and Hawaii that belong to role class 2 are mostly small airfields and industrial airports, which are embedded in local air route patterns.

IV. Conclusion

We show how directed flow patterns at all scales in directed networks can be harnessed using a combination of flow-based and structural approaches to uncover the different types of nodes that exist in a directed network. Both RBS and Markov Stability at their core rely on flows but each from a different stance: the former compares how the similarly nodes are positioned with respect to incoming and outgoing flows, while the latter establishes where flows tend to be trapped on a given timescale. The RSMT algorithm allows us to project complex datasets with local structure as true networks, facilitating its analysis with graph theoretical tools. Together, these techniques form a powerful framework for the analysis of directed networks which, as the examples here show, is applicable to networks originating from different disciplines.

Acknowledgments

MBD and MB acknowledge support from the UK EPSRC grant EP/I017267/1 under the Mathematics Underpinning the Digital Economy program. MBD acknowledges support from the James S. McDonnell Foundation Postdoctoral Program in Complexity Science/Complex Systems-Fellowship Award (#220020349-CS/PD Fellow). BV is supported through a PhD Award from the British Heart Foundation Centre of Research Excellence at Imperial College London (RE/08/002).

References

[1] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, pp. 39–43, 1953.
[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120.
[3] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.
[4] K. Cooper and M. Barahona, "Role-based similarity in directed networks," *arXiv:1012.2726*, Dec. 2010.
[5] K. Cooper, "Complex networks: Dynamics and similarity," Ph.D. dissertation, University of London, 2010.
[6] M. A. Carreira-Perpiñán and R. S. Zemel, "Proximity graphs for clustering and manifold learning," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 225–232.
[7] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, "Structural properties of the caenorhabditis elegans neuronal network."*PLoS Comput Biol*, vol. 7, no. 2, p. e1001066, 2011.
[8] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75 – 174, 2010.
[9] J.-C. Delvenne, S. Yaliraki, and M. Barahona, "Stability of graph communities across time scales," *P Nat Acad Sci USA*, vol. 107, no. 29, pp. 12 755–12760, 2010, also: arXiv:0812.1811 (2008).
[10] J.-C. Delvenne, M. T. Schaub, S. N. Yaliraki, and M. Barahona, "The stability of a graph partition: A dynamics-based framework for community detection," in *Dynamics On and Of Complex Networks, Volume 2*, A. Mukherjee, M. Choudhury, F. Peruani, N. Ganguly, and B. Mitra, Eds. Springer New York, 2013, pp. 221–242.
[11] R. Lambiotte, J. Delvenne, and M. Barahona, "Laplacian dynamics and multiscale modular structure in networks," *arXiv:0812.1770*, Dec. 2008.
[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J Stat Mech-Theory E*, vol. 2008, no. 10, p. P10008, 2008.
[13] M. Meilă, "Comparing clusterings—an information based distance," *J Multivariate Anal*, vol. 98, no. 5, pp. 873 – 895, 2007.
[14] T. Opsahl. (2011) Why anchorage is not (that) important: Binary ties and sample selection. [Online]. Available: http://wp.me/poFcY-Vw
[15] J. Kunegis, "KONECT – the Koblenz Network Collection," 2012. [Online]. Available: http://konect.uni-koblenz.de