DIGRAC: Digraph Clustering Based on Flow Imbalance

Yixuan He[1], Gesine Reinert[1,2], Mihai Cucuringu[1,2]

[1]University of Oxford, [2]The Alan Turing Institute

Motivation

- Most existing methods that could be applied to directed clustering use local edge densities as main signal and directionality as additional signal.
- We argue that even in the absence of any edge density differences, directionality can play a vital role in directed clustering as it can reveal latent properties of network flows.
- Therefore, instead of finding relatively dense groups of nodes in digraphs which have a relatively small amount of flow between the groups, our main goal is to recover clusters with **strongly imbalanced flow** among them, in the spirit of [1], where directionality (i.e, edge orientation) is the main signal.

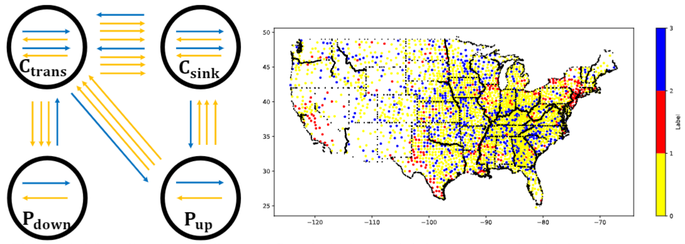


Figure 1: Visualization of directed flow imbalance:(a) A meta-graph which we hypothesize to be present on *Telegram* [2]. More edge weights flow from $\mathcal{C}_{\text{trans}}$ to $\mathcal{C}_{\text{sink}}$ than in the other direction; (b) top pair imbalanced flow on *Migration* data [3], along with the geographic locations of the counties and state boundaries (in black): most edges flow from red (1) to blue (2).

Problem Definition

We consider a directed graph (digraph) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with:

- The node set $\mathcal{V}$ is a set of n nodes.
- The edge set $\mathcal{E}$.
- The node feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times din}$.
- The adjacency matrix is denoted as $\boldsymbol{A}$ and is usually *asymmetric*.
- A **clustering** into K clusters: a partition of the node set into disjoint sets $\mathcal{V} = \mathcal{C}_0 \cup \cdots \cup \mathcal{C}_{K-1}$.
- *Self-supervised*: no label supervision during training.

Our DIGRAC Approach

- We introduce a graph neural network framework to obtain node embeddings for directed networks in a self-supervised manner, including a novel probabilistic imbalance loss for node clustering.
- We propose **directed flow imbalance** measures, which are tightly related to directionality, to reveal clusters in the network even when there is no density difference between clusters.

Probabilistic Cut

- Our **self-supervised** loss function is inspired by [1], aiming to cluster the nodes by maximizing a normalized form of cut imbalance across clusters.
- For K clusters, the *assignment probability matrix* $\boldsymbol{P} \in \mathbb{R}^{n \times K}$ has as row i the probability vector $\boldsymbol{P}_{(i,:)} \in \mathbb{R}^K$ with entries denoting the probabilities of each node to belong to each cluster; its k-th column is denoted by $\boldsymbol{P}_{(:,k)}$.
- The *probabilistic cut* from cluster $\mathcal{C}_k$ to $\mathcal{C}_l$ is defined as

$$W(\mathcal{C}_k, \mathcal{C}_l) = \sum_{i,j \in \{1,\ldots,n\}} \mathbf{A}_{i,j} \cdot \mathbf{P}_{i,k} \cdot \mathbf{P}_{j,l} = (\mathbf{P}_{(:,k)})^T \mathbf{A} \mathbf{P}_{(:,l)}.$$

Imbalance Flow and Probabilistic Volume

- The *imbalance flow* between clusters $\mathcal{C}_k$ and $\mathcal{C}_l$ is defined as
  $|W(\mathcal{C}_k, \mathcal{C}_l) - W(\mathcal{C}_l, \mathcal{C}_k)|, \quad \forall k, l \in \{0, \ldots, K-1\}.$
- For interpretability and ease of comparison, we normalize the imbalance flows to obtain an imbalance score with values in [0,1].
- The *probabilistic volume* for cluster $\mathcal{C}_k$ is defined as

$$VOL(\mathcal{C}_k) = VOL^{(\text{out})}(\mathcal{C}_k) + VOL^{(\text{in})}(\mathcal{C}_k) = \sum_{i,j} (\mathbf{A}_{i,j} + \mathbf{A}_{j,i}) \cdot \mathbf{P}_{j,k}.$$

$$\text{Then } VOL(\mathcal{C}_k) \geq W(\mathcal{C}_k, \mathcal{C}_l) \text{ for all } l \in \{0, \ldots, K-1\} \text{ and}$$

$$\min(VOL(\mathcal{C}_k), VOL(\mathcal{C}_l)) \geq |W(\mathcal{C}_k, \mathcal{C}_l) - W(\mathcal{C}_l, \mathcal{C}_k)|. \quad (1)$$

Imbalance Score

- The imbalance term, which is used in most of our experiments, denoted $\text{CI}^{\text{vol\_sum}}$, is defined as $\text{CI}^{\text{vol\_sum}}(k,l) = 2\dfrac{|W(\mathcal{C}_k, \mathcal{C}_l) - W(\mathcal{C}_l, \mathcal{C}_k)|}{VOL(\mathcal{C}_k) + VOL(\mathcal{C}_l)} \in [0,1]. \quad (2)$
- The aim is to find a partition that maximizes the imbalance flow under the constraint that the partition has at least two sets, to capture groups of nodes that could be viewed as representing clusters in the meta-graph. The normalization by the volumes penalizes partitions that put most nodes into a single cluster. The range [0,1] follows from Eq. (1).

Global Imbalance Objective

- To obtain a **global probabilistic imbalance score**, based on $\text{CI}^{\text{vol\_sum}}$ from Eq. (2), we average over pairwise imbalance scores of different pairs of clusters. Since the scores discussed are symmetric and the cut difference before taking absolute value is skew-symmetric, we only need to consider the pairs $T = \{(\mathcal{C}_k, \mathcal{C}_l) : 0 \leq k < l \leq K-1, k, l \in \mathbb{Z}\}$. We consider a ``*sort*'' variant to select these pairs. With $T(\beta) = \{(\mathcal{C}_k, \mathcal{C}_l) : (\mathcal{C}_k, \mathcal{C}_l) \in T, \text{CI}^{\text{vol\_sum}}(k,l)$ is among the top β values$\}$, we set

$$\mathcal{O}^{\text{sort}}_{\text{vol\_sum}} = \frac{1}{\beta} \sum_{(\mathcal{C}_k, \mathcal{C}_l) \in T(\beta)} \text{CI}^{\text{vol\_sum}}(k,l), \quad \text{and} \quad \mathcal{L}^{\text{sort}}_{\text{vol\_sum}} = 1 - \mathcal{O}^{\text{sort}}_{\text{vol\_sum}} \quad (3)$$

as the corresponding loss function.

Framework Overview

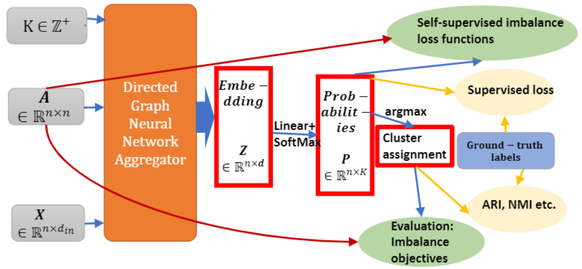
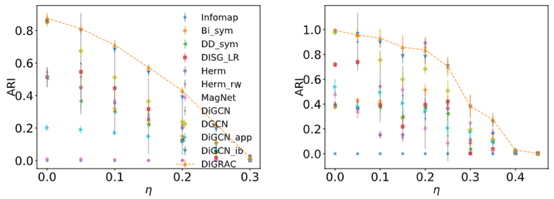


Figure 2: DIGRAC overview: from feature matrix $\boldsymbol{X}$, adjacency matrix $\boldsymbol{A}$, and number of clusters K, we first apply a directed GNN aggregator to obtain the node embedding matrix $\boldsymbol{Z}$, then apply a linear layer followed by a unit softmax function to get the probability matrix $\boldsymbol{P}$. Applying *argmax* on each row of $\boldsymbol{P}$ yields node cluster assignments. Green circles involve our proposed imbalance objective, while the yellow circles can only be used when ground-truth labels are provided.

DSBM Results



(a) DSBM("cycle", $n = 5000, p = 0.01$)

(b) DSBM("complete", $n = 1000, p = 0.1$)

Figure 4: Test ARI comparison on DSBMs, averaged over 50 runs. Dashed lines highlight DIGRAC's performance. Error bars indicate one standard error. These are DSBM results with $K = 5$ clusters and size ratio $\rho = 1.5$. Both networks contain ambient nodes, and p is the edge density.

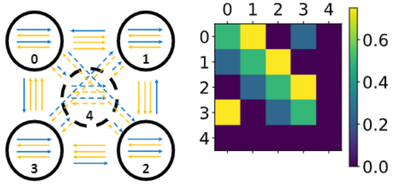
Real-World Data Results

Performance comparison on real-world data sets. The best is marked in **bold red** and the second best is marked in underline blue.

Metric	Data set	InfoMap	Bi_sym	DD_sym	DISG_LR	Herm	Herm_rw	DIGRAC
$\mathcal{O}^{\text{sort}}_{\text{vol\_sum}}$	*Telegram*	0.04±0.00	0.21±0.0	0.21±0.0	0.21±0.01	0.2±0.01	0.14±0.0	**0.32±0.01**
	Blog	0.07±0.00	0.07±0.0	0.0±0.0	0.05±0.0	0.37±0.0	0.0±0.0	**0.44±0.0**
	Migration	N/A	0.03±0.00	0.01±0.00	0.02±0.00	0.04±0.00	0.02±0.00	**0.05±0.00**
	WikiTalk	N/A	N/A	N/A	0.18±0.03	0.15±0.02	0.0±0.0	**0.24±0.05**
	Lead-Lag	N/A	0.07±0.01	0.07±0.01	0.07±0.01	0.07±0.02	0.07±0.02	**0.15±0.03**
$\mathcal{O}^{\text{naive}}_{\text{vol\_sum}}$	*Telegram*	0.01±0.00	0.26±0.0	0.26±0.0	0.26±0.01	0.25±0.02	0.23±0.0	**0.27±0.01**
	Blog	0.00±0.00	0.07±0.0	0.0±0.0	0.05±0.0	0.37±0.0	0.0±0.0	**0.44±0.0**
	Migration	N/A	0.01±0.00	0.01±0.00	0.01±0.00	0.02±0.00	0.01±0.00	**0.04±0.01**
	WikiTalk	N/A	N/A	N/A	0.1±0.02	0.04±0.0	0.0±0.0	**0.12±0.01**
	Lead-Lag	N/A	0.30±0.06	0.28±0.06	0.27±0.06	0.29±0.05	0.29±0.05	**0.32±0.11**

Directed Clustering Data

- Directed Stochastic Block Models (DSBMs).



(a) Meta-graph.

(b) Meta-graph adjacency matrix.

Figure 3: Visualization of a DSBM with a cycle meta-graph with ambient nodes, for a total of 5 clusters. 75% of the edges flow in the direction 0→1→2→3→0, while 25% flow in the opposite direction. Cluster 4 is the ambient cluster. In (a), dashed lines indicate flows with random equally likely directions; these flows do not exist in the meta-graph adjacency matrix. For (b), the lighter the color, the stronger the flow.

- Real-world data sets:

data set	n	$\lvert\mathcal{E}\rvert$	density	weighted	$\lvert\mathcal{E}^r\rvert$	$\frac{\lvert\mathcal{E}^r\rvert}{\lvert\mathcal{E}\rvert}(\%)$
Telegram	245	8,912	$1.28 \cdot 10^{-2}$	True	1,572	17.64
Blog	1,222	19,024	$1.49 \cdot 10^{-1}$	True	4,617	24.27
Migration	3,075	721,432	$7.63 \cdot 10^{-2}$	True	351,100	48.67
WikiTalk	2,388,953	5,018,445	$8.79 \cdot 10^{-7}$	False	723,526	14.42
Lead-Lag	269	29,159	$4.04 \cdot 10^{-1}$	True	0.00	0.00

References

[1] Cucuringu, M., Li, H., Sun, H., & Zanetti, L. (2020, June). Hermitian matrices for clustering directed graphs: insights and applications. In International Conference on Artificial Intelligence and Statistics (pp. 983-992). PMLR.

[2] Bovet, A., & Grindrod, P. (2020). The activity of the far right on telegram.

[3] Perry, M. J. (2003). State-to-state Migration Flows, 1995 to 2000. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.