GNNs for Node Clustering in Signed and Directed Networks

Yixuan He
yixuan.he@stats.ox.ac.uk
University of Oxford
Oxford, UK

ABSTRACT

With an increasing number of applications where data can be represented as graphs, graph neural networks are a useful tool to apply deep learning to graph data. In particular, node clustering is an important problem in network analysis. Signed and directed networks are important types of networks that are linked to many real-world problems; their asymmetry provides a challenge for many clustering methods.

We propose two graph neural network models for node clustering in signed networks and directed networks, respectively. The methods are end-to-end in combining embedding generation and clustering without an intermediate step. Experimental results on a synthetic signed stochastic block model, a polarized version of it, and real-world data at different scales, demonstrate that our proposed methods can achieve comparable or better results than state-of-the-art node clustering methods, for a wide range of noise and sparsity levels. The introduced models complement existing well-performing methods through the possibility of including exogenous information, in the form of node-level features or labels.

edges has +/-
labels

CCS CONCEPTS

• **Information systems** → **Clustering**.

KEYWORDS

graph neural networks, signed networks, directed networks

ACM Reference Format:
Yixuan He. 2022. GNNs for Node Clustering in Signed and Directed Networks. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM ’22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3488560.3502212

1 MOTIVATION

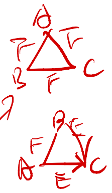
With an increasing number of applications where data from non-Euclidean domains are represented as graphs, (e.g., social networks, citation networks, and biochemical graphs), graph data, which contains rich relation information, are related to many learning tasks ; clustering is one of these. Graph data are usually not i.i.d.; graph neural networks (GNNs) have been developed to extend standard deep learning tools to graph data [11].

measure the quality of a partition

high dense within communities,
sparse connection between different.

Related to clustering in networks is the task of community detection, whose goal is to partition the node set of a network such that, loosely speaking, nodes within a cluster should be similar to each other, while nodes across clusters should be dissimilar [12]. The quality of a partition is often assessed through a modularity objective function which compares the partition to that expected under a null model for the network with the assumption that nodes within a cluster are relatively more densely connected than nodes across clusters. However, depending on the task at hand, *similarity could have different meanings*. In a signed network with positive and negative edges, similarity may relate to the neighbourhood of a node such as the proportion of shared friends or enemies. In a directed network, nodes could also be clustered depending on their position within a directed flow on the network, see [4].

Signed Graphs The main novelty of our approach for signed clustering is a new take on the role of social balance theory for signed network embeddings [5]. The standard heuristic for justifying the criteria for the embeddings hinges on the assumption that in a social network, *balanced* triangles are preferred; these are triangles such that either all three nodes are friends, or two friends have a common enemy; otherwise it would be viewed as *unbalanced*. More generally, all cycles are assumed to prefer to contain either zero or an even number of negative edges. This hypothesis is supported empirically for unsigned friendship networks, but is difficult to justify for general signed networks. For example, the relationship between trust and distrust may not be a simple negation; the enemies of enemies are not necessarily friends; an example is the social network of relations between 16 tribes of the Eastern Central Highlands of New Guinea [7]. Hence, we take a neutral stance on whether an enemy’s enemy is a friend, see Figure 1.



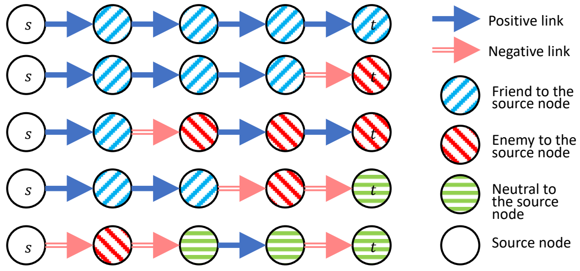
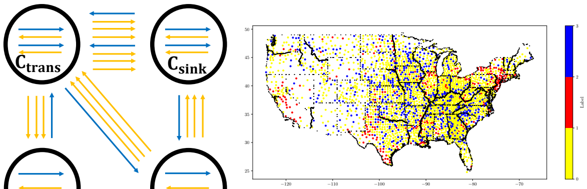


Figure 1: Example: five paths between the source (s) and target (t) nodes, and resulting relationships. While we assume a neutral relationship on the last two paths, social balance theory claims them as "friend" and "enemy", respectively.

Directed Graphs. The main novelty for our directed clustering approach is an objective based on flow imbalance. While 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

ACM ISBN 978-1-4503-9132-0/22/02.
https://doi.org/10.1145/3488560.3502212



(a) A meta-graph which we hypothesize to be present on *Telegram*.

(b) Strongest imbalanced flow on *Migration* data, along with the geographic locations of the counties and state boundaries (in black).

Figure 2: Visualization of directed flow imbalance: (a) there are much more edge weights flow from C_{trans} to C_{sink} than the other direction; (b) top pair imbalanced flow on *Migration* data [6]: most edges flow from red (1) to blue (2).

Directionality

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 *strong and imbalanced flow* among them, in the spirit of [2], where directionality (i.e, edge orientation) is the main signal. In contrast to standard approaches focusing on edge density, here edge directionality is not a nuisance but the main piece of information to uncover the latent structure [4]. Figure 2(a) plots a meta-graph which we hypothesize to be present for *Telegram* [1], where most edge weights flow from the core-transient cluster (C_{trans}) to the core-sink cluster (C_{sink}) than the other direction. As another real-world example, Figure 2(b) shows the strongest flow imbalances between clusters in a network of US migration flow [6]; most edges flow from the red cluster (1) to the blue one (2).

2 BACKGROUND AND RELATED WORK

We introduce GNN methods for node clustering, using the idea of an aggregator. Here we shall use powers of adjacency matrices for aggregation which is motivated by [9].

For *signed clustering*, while GNNs have been utilized for signed network embedding tasks – for example, SGCN [3] utilizes social balance theory to aggregate and propagate the information across layers – they have not yet been employed for signed clustering.

For *directed clustering*, [2] seeks to uncover clusters characterized by a strongly imbalanced flow circulating among them, based on eigenvectors of a Hermitian matrix. [13] introduces a complex Hermitian matrix that encodes undirected geometric structure in the magnitude of its entries, and directional information in their phase. DiGCN [10] builds a directed Laplacian based on PageRank, and aggregates information dependent on higher-order proximity. InfoMap [8] assumes that there is a "map" underlying the network, similar to a meta-graph in our DIGRAC approach [4].

Square matrix
↓
equals its
conjugate
transpose
$A = A^H$
eigenvectors
?

3 METHODOLOGY AND EXPERIMENTS

Most state-of-the-art methods generating node embeddings of signed networks focus on link sign prediction, and those that pertain to node clustering are usually not GNN methods. In [5], we introduce a novel probabilistic balanced normalized cut loss for training nodes in a GNN framework for semi-supervised signed network clustering, called SSSNET. The main novelty of our approach is a new take on the role of social balance theory for signed network embeddings. Figure 1 illustrates five different paths of length four, connecting the source and the target nodes. We can also obtain the relationship of a source node to a target node within a path by reversing the arrows in Figure 1. Note that it is possible for a node to be both a "friend" and an "enemy" to a source node simultaneously, as there might be multiple paths between them, with different resulting relationships. Our model aggregates these relationships by assigning different weights to different paths connecting two nodes. For example, the source node and target node may have all five paths shown in Figure 1 connecting them. Since the last two paths are neutral paths and do not cast a vote on their relationship, we only take the top three paths into account.

For directed networks, 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. In [4], 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. Figure 2(a) gives an example of a meta-graph.

make prediction
in supervised manner
without explicit
labels

Directed
↓ GNN
Node Embedding
Proba Normalized
Cut loss

To train a GNN model, we devise differentiable loss functions. As cluster assignment outputs can be probabilistic, a probabilistic version of the balanced normalized cut loss is proposed in [5], while a probabilistic version of cut flow imbalance loss is introduced in [4]. As evaluation methods for node clustering, the Adjusted Rand Index (ARI) is used. Depending on the downstream task, other measures are employed, such as the *unhappy ratio* by [5], and the *imbalance scores* by [4]. As for data sets to validate our proposed methods, we generate representative synthetic data, in the form of a signed or directed stochastic block model, with possibly unequal cluster sizes but equal edge density, and with a polarized signed stochastic block model for signed clustering. At the same time, we test our models' efficacy on real-world data sets. Empirical results show that our GNNs give comparable and usually better results than state-of-the-art node clustering methods.

↓
Directionality
Imbalanced
flow

Future work will extend our work to temporal networks. It will include building packages for the signed and directed networks so as to facilitate research in the field.

REFERENCES

[1] A. Bovet and P. Grindrod. 2021. The Activity of the Far Right on Telegram. *Complex Networks and its Applications.*

[2] M. Cucuringu et al. 2020. Hermitian matrices for clustering directed graphs: insights and applications. In *AISTATS.*

[3] T. Derr et al. 2018. Signed graph convolutional networks. In *ICDM*. IEEE.

[4] Y. He et al. 2021. DIGRAC: Digraph Clustering with Flow Imbalance. *arXiv preprint arXiv:2106.05194* (2021).

[5] Y. He et al. 2021. SSSNET: Semi-Supervised Signed Network Clustering. *arXiv preprint arXiv:2110.06623* (2021).

[6] M. Perry. 2003. *State-to-state Migration Flows, 1995 to 2000.* US Department of Commerce, Economics and Statistics Administration, US.

[7] K. Read. 1954. Cultures of the central highlands, New Guinea. *Southwestern Journal of Anthropology* 10, 1 (1954), 1–43.

[8] M. Rosvall and C. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proc. of the Nat. Acad. of Sciences* 105, 4 (2008).

[9] Y. Tian et al. 2019. Rethinking kernel methods for node representation learning on graphs. *NeurIPS* (2019).

[10] Z. Tong et al. 2020. Digraph Inception Convolutional Networks. *NeurIPS* (2020).

[11] Z. Wu et al. [n. d.]. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* ([n. d.]), 1–21.

[12] J. Zhang et al. 2021. Directed community detection with network embedding. *J. Amer. Statist. Assoc.* (2021), 1–11.

[13] X. Zhang et al. 2021. MagNet: A Neural Network for Directed Graphs. *NeurIPS* (2021).