Survey of Community Detection Approaches for Signed Graph Networks

Maria Tomasso, met48@txstate.edu, https://mtomasso.github.io
Advisor: Dr. Jelena Tešić, jtesic@txstate.edu
Department of Computer Science, Texas State University, San Marcos, TX 78666

Introduction

Goal: Community structure(s) discovery in large signed and weighted graph networks at scale.
Approach: apply unsigned graph clustering to signed graphs by **adapting unsigned** clustering methods and building upon existing **signed graph clustering** approaches.
Challenges: *L*oss of signed edge information using traditional unsigned methods; Assumptions for unsigned clustering e.g. semi-definite Laplacian for spectral clustering does not hold for signed graphs in most cases.

Signed Graphs and Balance

A **Graph** $\mathcal{G}$ consists of two disjoints sets, a set of *vertices* v , $v \in \mathcal{V}$ and a set of *edges* e, $e \in \mathcal{E}$. Graphs can be **directed** or **undirected**. Signed graph edges are assigned $+1$ or -1 weights. **Complete** graph has every pair of vertices connected by an edge. **Dense** graph is a graph with number of edges close to the maximum number of edges. **Sparse** graph has very few edges.

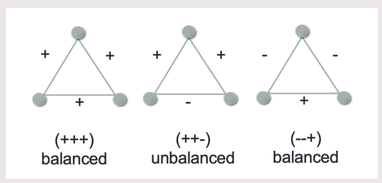

Figure: Balance Graph Theory example on the triangle signed graph

Balance Theory is a sociological Mathematical approach introduced by Fritz Heider in 1946. It provides a framework for determining when a social network is stable, and is based in study of human perception: "The enemy of my enemy is my friend".

Graphs as Matrices

For a sample graph $\mathcal{G}$ with 4 nodes and 5 edges, the **adjacency matrix** $\mathcal{A}$ is a square matrix whose row and column numbers represent vertices. Entries in the adjacency matrix indicate the edges.

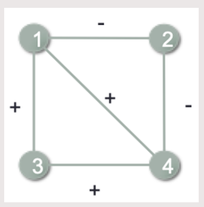

Figure: Sample graph $\mathcal{G}$

$$A = \begin{bmatrix} 0 & -1 & 1 & 1 \\ -1 & 0 & 0 & -1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & 1 & 0 \end{bmatrix}$$

Figure: **adjacency matrix** $\mathcal{A}$ is a square matrix where *(i,j)* entry is an edge between nodes *i* and *j*

Tthe **degree matrix** $\mathcal{D}$ is a square matrix whose row and column numbers represent vertices. Entries along the diagonal represent the degrees of the vertices; all other entries are zero. The **Laplacian matrix** $\mathcal{L}$ is equal to the degree matrix $\mathcal{D}$ minus the adjacency matrix $\mathcal{A}$.

$$D = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

Figure: The **degree matrix** $\mathcal{D}$:diagonal entry at *(i,i)* is the degree of the vertex *i*

$$L = \begin{bmatrix} 3 & 1 & -1 & -1 \\ 1 & 2 & 0 & 1 \\ -1 & 0 & 2 & -1 \\ -1 & 1 & -1 & 3 \end{bmatrix}$$

Figure: The **Laplacian matrix** $\mathcal{L} = \mathcal{D} - \mathcal{A}$

Signed Graph Analysis

Node-based: **community detection**, node ranking, node classification, node embedding
Link-oriented: Link prediction, sign prediction, tie strength prediction, negative link prediction
Application oriented: Information diffusion, recommendation, data classification, data clustering

Year	Event
1946	Heider publishes landmark paper 'Attitudes and cognitive organization', introducing the idea of structural balance theory
1956	Cartwright and Harary built on Heider's work and formalized the idea of a balanced signed graph
1978/ 1979	Moore applied structural balance theory to diplomatic relations in South Asia
1993	Axelrod and Bennet applied structural balance theory to international relations during WWII
2005	Massa and Avesani applied signed graph theory to the Slashdot Zoo network
2005+	Signed graph mining exploded in popularity with the rise of social media

Figure: Timeline of the signed graph analysis and mining

Adapted Unsigned Clustering

Adapted Spectral Clustering

- These methods involve adapting the original spectral clustering methodology proposed by Shi and Malik in 2000
- The goal is to **modify either the graph or the Laplacian to ensure all eigenvalues are real and non-negative**
- Many incremental improvements have been introduced in the past decade

Kunegis 2009 (**Signed Laplacian**)
- Positive definite iff the network is unbalanced
- Separates negative-linked nodes rather than grouping positive-linked nodes

Hseih 2012 (**Clustering with matrix completion**)
- Relies heavily on sign inference
- Complete the graph with any matrix completion algorithm, then perform spectral clustering

Chiang 2012 (**Balanced normalized cut**)
- Minimizes positive links between communities

Zheng 2015 (**Balanced normalized signed Laplacian**)
- Addresses shortcomings of the signed Laplacian

Mercado 2016 (**Geometric Laplacian means**)
- Uses the geometric mean of the positive and negative Laplacians
- This approach works well for networks with little to no noise in either the positive or negative structure
- Effective, but computationally expensive

Mercado 2019 (**Matrix power means**)
- Uses the matrix power mean of the normalized and signless positive and negative Laplacians

Other Spectral Methods

- These methods do **not** modify traditional spectral clustering; however, they utilize eigenvalues

Cucuringu 2019 (**SPONGE**)
- Solves for the k smallest eigenvalues using a generalized eigenproblem

Synthetic Example

- Three synthetic graphs were generated using signed stochastic blockmodels
- The Rand index measures the similarity between two clusterings and varies from 0 to 1
- Clusters were generated using the signed Laplacian, balanced normalized cut, and SPONGE for each graph

Example 1 - 100 nodes, 20 communities

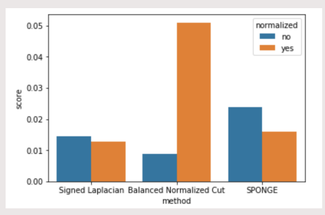

Figure: Comparison of Rand indices for signed spectral clustering, balanced normalized cuts, and SPONGE clustering on example 1

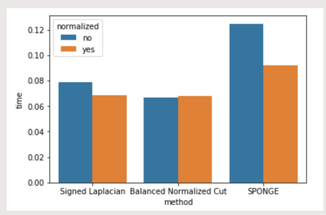

Figure: Comparison of runtimes for signed spectral clustering, balanced normalized cuts, and SPONGE clustering on example 1

Example 2 - 5000 nodes, 50 communities

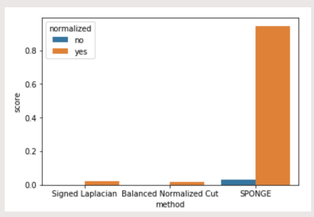

Figure: Comparison of Rand indices for signed spectral clustering, balanced normalized cuts, and SPONGE clustering on example 2

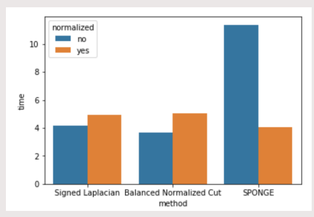

Figure: Comparison of runtimes for signed spectral clustering, balanced normalized cuts, and SPONGE clustering on example 1

Example 3 - 10,000 nodes, 100 communities

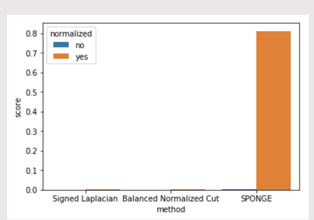

Figure: Comparison of Rand indices for signed spectral clustering, balanced normalized cuts, and SPONGE clustering on example 3

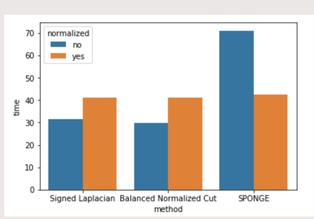

Figure: Comparison of runtimes for signed spectral clustering, balanced normalized cuts, and SPONGE clustering on example 3

Signed Graph Clustering

Early Work

- Early work in signed graphs focused on smaller, denser networks
- These algorithms worked well at the time, but are not as effective on larger, sparser graphs

Brieger, 1975 (**CONCOR**)
- Repeatedly partitions nodes into two sets based on convergence of an iterative correlation matrix

White, 1976 (**Blockmodels**)
- Introduced a method for identifying social structure when multiple types of ties are present

Modularity-Based Methods

Main idea: Seek densely connected clusters using optimization techniques

Newman 2004
- Introduced an hierarchical agglomeration algorithm based on modularity for community detection
- Notable for running in near linear time – a significant improvement over state-of-the-art techniques at the time

Yang 2007 (**FEC**)
- Designed for densely connected networks
- Treats sign and density of edges as clustering attributes
- Designed for signed graphs but applicable to unsigned graphs

Gomez 2009
- Extended existing methods to networks that are signed, directed, weighted, and/or contain loops

Anchuri 2012
- Uses an iterative approach to minimize frustration and maximize modularity

Traag 2013
- Uses subgraph analysis to determine appropriate resolution parameters

Random Walk Models

Main idea: Trace a random path from node to node on a graph, with edge weights affecting the probability of a node being included in the walk

Harel 2001
- Introduced random walk clustering for weighted graphs

Hua 2020
- Proposes the random walk gap (RWG) heuristic for clustering on signed graphs
- Uses two types of random walks: one on positive edges, and one on negative edges, to make inferences about community structure
- First study to use walks on negative edges

Conclusion and Outlook

- Community detection in signed graphs is a rapidly evolving discipline with many different underlying methodologies
- Understanding the underlying assumptions and methodologies is critical when choosing the correct algorithm for your data
- While spectral methods remain as the "gold-standard", **emerging methodologies provide promise in community detection as well as other areas of signed graph mining**
- See https://datalab12.github.io/work/attitudinal-network-science.html for more information on a new methodology being developed here at Texas State!

References

- See https://mtomasso.github.io for full citation list
- Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of Heider's theory. Psychological review 63, 5 (1956), 277.
- M Cucuringi, P Davies, A Glielmo, H Tyagi (2019). SPONGE: A generalized eigenproblem for clustering signed networks. arXiv preprint arXiv:1904.08575
- Tang, Jiliang, Yi, Chang, and Huan, Liu (2016). A Survey of Signed Network Mining in Social Media. arXiv preprint arXiv1511.07569

DataLab12.github.io