Abstract

Community detection is a common task in social network analysis (SNA) with applications in a variety of fields including medicine, criminology, and business. Despite the popularity of community detection, there is no clear consensus on the most effective methodology for signed networks. In this paper, we summarize the development of community detection in signed networks and evaluate current state-of-the-art techniques on several real-world data sets. First, we give a comprehensive background of community detection in signed graphs. Second, we compare various adaptations of the Laplacian matrix in recovering ground-truth community labels via spectral clustering in small signed graph data sets. Third, we evaluate the scalability of leading algorithms on the Wikipedia Elections and Correlates of War data sets. Finally, we make recommendations for extensions and improvements to existing methodologies.

Community Discovery in Signed Networks: A Survey and Evaluation of Existing Methodologies

Maria Tomasso - CS 7387 Spring 2021

May 14, 2021

1 Introduction

With the rise of social media and the availability of large signed data sets we have seen increased interest in data mining methodologies for signed graphs. One especially useful sub-task of signed graph data mining is community detection. In its most general terms, a community within a network is defined as a partitioning of nodes such that nodes within the same cluster are strongly connected while nodes in different clusters are weakly connected; in essence, similar nodes should be grouped together. In real data sets, community structure is almost always present to some degree [8].

Identifying and studying communities within networks has applications in a variety of fields. Examples of phenomena discoverable through community detection include bot activity and fraud in criminology, customer segmentation in marketing, astroturfing in political science, and tumor detection and environmental hazards in public health [13]. Additionally, understanding a network's community structure can help researchers understand the propagation patterns of vectors spreading through the network, including contagious diseases and fake news [8]. The above examples were all originally studied in unsigned networks, but researchers have recently turned their attention to the added information provided by the presence of negative links in networks [7]. While community detection in unsigned networks relies on the absence of connections between a pair of nodes to determine if they belong in different communities, the presence of negative links between provides affirmative evidence of their dissimilarity. Thus, signed networks provide a unique opportunity to achieve stronger community distinctions when the presence of negative edges is considered. Furthermore, when negative edges are included in a network additional applications become available. For example, signed networks are particularly useful when studying social dynamics and stability with respect to friendship and enmity [1] and more recently have been used to study the behavior of the brain [24].

In this paper we will first present an overview of the work that has been done on community detection in signed networks to date. The methods are divided roughly into three categories: methods adapted from unsigned methods; methods that are agnostic to signed vs unsigned; and methods that work only for signed graphs. Further subcategories within each subcategory are also defined. Second, we will compare current state-of-the-art methodologies on small signed networks with known ground-truth communities and compare their ability to recover the ground-truth labels based on edge signs alone. Third, we will evaluate the scalability of leading clustering methodologies on attitudinal data sets from Wikipedia Elections [16], Slashdot [16], and the Correlates of War [25] data sets.

2 Background

The study of signed graphs began in the 1940s with the seminal work of sociologist Fritz Heider on social balance theory [11]. In 1956, Cartwright and Harary took the notion of social balance and applied it to signed graphs, initiating a new branch of graph theory focused on the study of balance in graphs [10]. Researchers quickly realized that signed graphs were a perfect vehicle for modeling expressed opinions between entities through edge sign, known as attitudinal network graphs. Early applications of balance in attitudinal network graphs included modeling diplomatic relations in the Middle East [19], South Asia [20], and and Allied and Axis powers during World War II [2].

Before discussing clustering methodologies, we will begin with some definitions. A **Graph** $\mathcal{G}$ consists of two disjoints sets: a set of *vertices* v ,$v \in \mathcal{V}$ and a set of *edges* e, $e \in \mathcal{E}$. In this paper, graph and network can be assumed to be synonymous. Graphs can be **directed** or **undirected**. In a directed graph, an edge may connect node i to node j without node j necessarily being connected to node i. In an undirected graph, if node i is connected to node j then node j must be connected to node i. A graph can be **weighted**, which means that each edge has a 'weight' attribute that can represent the strength of the connection. In a signed graph, edges are assigned +1 or -1 weights. Since signed graphs can be used to model opinions, they are also called **sentiment networks** or **attitudinal networks**. In a **complete** graph every pair of vertices is connected by an edge. A **dense** graph is a graph with number of edges close to the maximum number of edges, while a **sparse** graph has very few edges relative to the number of nodes. Most modern graph data sets of interest are large and sparse, i.e. social media networks where most users are only connected to a small fraction of the overall community, but smaller and more specialized networks also occur in the real world. The Highland Tribes [22] and Sampson (citation needed) data sets are two examples of real-world signed and dense graphs that will be evaluated in this paper.

Community detection began with unsigned graphs and has been extended to signed graphs in recent years. In graph theory, a **community** is defined as a cluster of nodes such that nodes within the cluster are densely connected to other nodes within the cluster and sparsely connected to those outside of the cluster. Two algorithms, **spectral clustering** and **kernel k-means clustering**, form the basis for most community detection algorithms for unsigned graphs. To extend the concept of a community to signed graphs, we seek clusters such that nodes are positively connected to nodes within their own cluster and negatively connected those those outside of the cluster. Theorems based on **social balance theory**, a branch of signed graph theory proposed by [11] and developed by [10] in the 1940s and 1950s, allows certain well-behaved graphs to be 'perfectly' partitioned such that negative edges exist only between clusters and positive edges exist only within clusters. In the real world, however, such well-behaved data sets are uncommon and other methods based on discrete optimization must be used. This is where signed graph clustering methodologies find their use. Since many signed graph clustering methodologies were adapted directly from those designed for unsigned graphs, we will begin with a brief overview of clustering in unsigned graphs.

2.1 Spectral clustering

Spectral clustering was introduced by Shi and Malik [26] in 2000 and aims to group nodes into disjoint partitions such that nodes is the same partition are 'similar' and nodes in separate partitions are dissimilar'. The two central centrals of spectral clustering are (1) What criterion determines if a partition is 'good', and (2) how can partitions fitting the above criteria be efficiently computed. Shi and Malik introduced the **normalized cut** as a criterion for a 'good' partition and showed that it can be formulated as a generalized eigenvalue problem. Since minimizing the normalized cut is NP-complete, the goal of spectral clustering is to find an approximate discrete solution efficiently.

2.2 Weighted kernel k-means clustering

Weighted kernel k-means clustering, on the other hand, improves traditional k-means clustering by applying a non-linear mapping from the original data to a higher dimensional space using a **kernel function**. This allows clusters that are non-linearly separable to be detected. We can further generalize the problem by assigning weights to each point. After assigning weights and applying the non-linear mapping, traditional k-means clustering can be run. Note that the selection of the kernel and weights is critical in this type of clustering [6].

Spectral clustering and weighted kernel k-means clustering are both highly effective on unsigned graphs, but neither can be directly extended to signed graphs without prior modifications. In this paper we explore methodologies for clustering signed graphs that have been derived from unsigned methods, as well as novel methodologies that are been developed exclusively for signed graphs.

2.3 Prior Surveys

In 2016, Tang et al. published a comprehensive survey of techniques used in signed graph mining including community detection. In this paper, we build on their work by delving deeper into the mathematics behind each clustering technique and running experiments to assess and compare leading methodologies on a variety of real-world networks. Furthermore, we discuss the 5 years of development in the field that have occurred since the publication of Tang's 2016 survey [27].

2.4 Literature Review

2.4.1 Reference Searching

In the first stage of the literature review, both forwards and backwards reference searching were used to find significant contributions to the field.

2.4.2 Systemic Review

After forwards and backwards reference searching, a systemic literature review was conducted to find any publications on signed graph clustering that were previously missed.

The systemic literature review was conducted using five databases: the ACM Digital Library (https://dl.acm.org), Cornell's arXiv (https://arxiv.org), IEEE Xplore (https://ieeexplore.ieee.org/Xplore/home.jsp), SpringerLink (https://link.springer.com), and CiteSeerX (https://citeseerx.ist.psu.edu/index). Search terms used were:

- ”signed” AND ”graph” AND ”clustering”
- ”signed” AND ”graph” AND ”community” AND ”detection”

All results were saved and manually reviewed for relevance.

3 Adaptations of methods developed for unsigned graphs

Spectral clustering has long been considered the gold standard in community detection for unsigned graphs. Introduced by Shi and Malik in 2000, this methods exploits the fact that, in an unsigned graph with n separate connected components, n eigenvalues of the graph's Laplacian will be zero. A graph with separate connected components represents the trivial problem in community detection. However, positive eigenvalues that are very close to zero represent densely connected components with few edges between them - the goal of most community detection algorithms. The original algorithm attempts to find a global minimal solution to the normalized cut, a measure that counts edges between clusters after normalizing by the total number of edges adjacent to the cluster. Spectral clustering begins by finding the Laplacian of the matrix representation of the network. Since Laplacians be be calculated in different ways, this introduces variations on the spectral clustering technique. After finding the Laplacian, the eigenvalues are computed. Eigenvalue computation is often expensive and prone to error for very large matrices, so if reasonable bounds on the problem are known (i.e. the maximum number of clusters) the problem can be reduced to finding the k smallest eigenvalues. After finding the eigenvalues, they are plotted in increasing order and the eigengap, the largest 'early' increase in sequential eigenvalues, is identified. The location of the eigengap indicates the proper number of clusters to use, and there is some ambiguity to this step depending on the context of the problem. After identifying the number of clusters, k-means can be applied to cluster the communities [26].

The original spectral clustering methodology has been adapted for use on signed graphs with the help of social balance theory. Before introducing the adapted spectral clustering methodologies, a basic understanding of balance theory is needed. Social balance theory was introduced by Fritz Heider in 1946 and mathematically formalized by Cartwright and Harary in 1956. If a graph is *balanced* or *weakly balanced* it can be said that an underlying community structure exists in the graph, thus social balance theory continues to form the foundation for many community detection algorithms for signed graphs to this day.

Theorem 3.1. *A network is* **balanced** *if and only if (i) all of its edges are positive, or (ii) the nodes can be partitioned into two distinct clusters such that all edges within a cluster are positive and all edges between clusters are negative. Such a partition is known as a* **Harary cut***.*

Theorem 3.2. *a network is* ***weakly balanced*** *if and only if (i) all of its edges are positive, or (ii) the nodes can be partitioned into k distinct clusters such that all edges within a cluster are positive and all edges between clusters are negative.*

For balanced graphs, we can apply a Harary cut initially for a 2-way cut and then apply unsigned spectral clustering [26] to each cluster for the k-way clustering. Thus, for a fully balanced graph the only modification to unsigned methods that is needed is an initial Harary cut. Additionally, for a weakly balanced graph we can begin by partitioning the nodes into n clusters with only positive edges within clusters and negative edges between clusters, then apply unsigned clustering to each individual cluster for the k-way clustering. Unfortunately, getting the initial n-way clustering for a weakly balanced graph is only computationally trivial if $n = 2$. Thus, more sophisticated techniques are required for graphs that are not balanced.

When dealing with real data sets, balanced graphs are rare. The following section will discuss modifications to traditional spectral clustering that are specifically for weakly balanced signed graphs.

The first measure we will introduce for assessing two-way clustering of a graph is the **graph cut**.

Definition 3.1. *For an unsigned graph G with disjoint clusters X and Y,* $cut(X,Y) = \sum_{i \in X j \in Y} A_{ij}$

The cut is essentially the number of edges or, for a weighted graph, the sum of weights between clusters. Since the typical goal of clustering is to group densely connected nodes together, choosing X and Y to minimize the cut is a good first step. Unfortunately, the cut does not account of the size of clusters and can therefore lead to separating few to single vertices if applied as-in. To remedy this, the **ratio cut** is introduced.

Definition 3.2. *For an unsigned graph G with disjoint clusters X and Y,* $ratiocut(X,Y) = cut(X,Y)(\frac{1}{|X|} + \frac{1}{|Y|})$

These measures must be adapted for use on a signed network.

Definition 3.3. *For a signed graph G, the* ***signed graph cut*** *is given by* $scut(G) = 2 * cut^{+}(X,Y) + cut^{-}(X,X) + cut^{-}(Y,Y)$*.*

Definition 3.4. *The signed ratio cut is given by* $SignedRatioCut(X,Y) = scut(X,Y)(\frac{1}{|X|} + \frac{1}{|Y|})$*.*

Definition 3.5. *The signed normalized cut is give by* $SignedNormalizedCut(X,Y) = scut(X,Y)(\frac{1}{vol(X)} + \frac{1}{vol(Y)})$*, where* $vol(X)$ *and* $vol(Y)$ *represent the sum of the degrees of the nodes in X and Y, respectively.*

While effective on some graphs, traditional spectral clustering assumes (1) that all eigenvalues of the Laplacian are non-negative, i.e. the Laplacian is positive semidefinite, and (2) that the eigenvalues of the Laplacian can be efficiently computed. This is where we run into two issues with real signed networks. First, we cannot guarantee that the Laplacian of a signed graph will be positive semi-definite if we compute it using the standard definition of $L = D - A$, the degree matrix minus the adjacency matrix. Second, real signed graph data sets are often very large which presents issues with common methods for solving for eigenvalues. Iterative algorithms are not guaranteed to converge for ill-posed matrices.Thus, any method seeking to adapt spectral clustering to signed graphs must ensure that (1) eigenvalues are real and non-negative and (2) the new procedure is scalable to large networks.

3.1 Laplacian positive semidefinite

In this section we will review the incremental advances in modifying spectral clustering for signed graphs through a lens of accuracy and scalability.

Definition 3.6. *The Laplacian matrix of an unsigned graph G is given by* $L = D - A$ *where D represents the diagonal degree matrix and A represents the adjacency matrix.*

The standard Laplacian matrix of a signed graph is indefinite and thus will not yield real, non-negative eigenvalues. Kunegis, et al defines a modified Laplacian matrix.

Definition 3.7. *[15] The signed Laplacian matrix (also called the combinatorial Laplacian) of a graph G is given by $\bar{L} = \bar{D} - A$, where $\bar{D}$ is the signed degree matrix given by $\bar{D}_{ii} = \sum_{j\sim i} |A_{ij}|$.*

The authors prove that the signed Laplacian is positive semidefinite and, in some cases, positive definite, thus guaranteeing this Laplacian is a suitable basis for spectral clustering. Spectral clustering using the signed Laplacian is shown to be equivalent to the k-way signed ratio cut problem, which counts positive edges between clusters and negative edges within clusters. Two additional normalizations of the signed Laplacian matrix are also proposed. These can also be used for clustering and this practice corresponds to signed normalized cuts.

Definition 3.8. *[15] The random walk normalized Laplacian for signed graphs is given by $\bar{L}_{rw} = I - \bar{D}^{-1}A$. $\bar{L}_{rw}$ is positive semi-definite.*

Definition 3.9. *[15] The symmetric normalized Laplacian for signed graphs (Kunegis 2009) is given by $\bar{L}_{sym} = \bar{D}^{-1/2}\bar{L}\bar{D}^{-1/2} = I - \bar{D}^{-1/2}A\bar{D}^{-1/2}$. It should be noted that normalized Laplacians tend to yield better results than unnormalized Laplacians for graphs with skewed degree distributions.*

Theorem 3.3. *[15] The signed Laplacian matrix of a graph is positive-definite if and only if the graph is unbalanced.*

Hseih et al impute edges between unconnected nodes assuming balance. After creating a fully connected, maximally balanced graph based on the initial data, the k smallest eigenvalues are computed before k-means clustering is run to obtain node labels.

Chiang et al addresses shortcomings in the signed Laplacian $\bar{L}$ that arise when extending the problem to a k-way clustering problem by introducing the *balance normalized cut*, a criterion for k-way clustering problems that is analogous to the normalized cut. Additionally, Chiang et al introduce a multilevel framework that refines results by first dividing nodes into levels, and then applying their modified version of spectral clustering to each level. This approach enhances scalability by applying clustering to subsets of nodes rather than the entire graph at once. This approach was shown to be highly scalable (and tested on graphs as large as 1 million nodes and 100 million edges with a run-time under 400 seconds) and comparable with other state-of-the-art approaches. [4]

Theorem 3.4. *[4] There does not exist any representation of $\{\mathbf{x_1}, ..., \mathbf{x_k}\}$ such that the general k-way signed ratio cut objective is minimized.*

The above theorem holds because the k-way signed ratio cut inherently has less available information about each node than the two way signed ratio cut when $k > 2$. If there are only two clusters, c_1 and c_2, and we know that node i and node j both do not belong to c_1, they clearly both belong to c_2 are are therefore in the same cluster. However, if $k > 2$, we can not infer that if two nodes are both excluded from one cluster they must share another cluster. Thus, without modification minimizing the k-way signed ratio cut will not yield an optimal solution. To remedy this, Chiang et al propose a series of new objectives that do not pose this problem when extended to k-way clustering.

Definition 3.10. *[4] The **positive ratio association** maximizes the number of positive edges within each cluster relative to the clusters size. It is given by $\max_{\{\mathbf{x_1},\ldots,\mathbf{x_k}\}\in I} \sum_{c=1}^{k} \frac{\mathbf{x_c^T} A^{+} \mathbf{x_c}}{\mathbf{x_c^T x_c}}$.*

Definition 3.11. *[4] The **negative ratio association** minimizes the number of negative edges within each cluster relative to the clusters size. It is given by $\min_{\{\mathbf{x_1},\ldots,\mathbf{x_k}\}\in I} \sum_{c=1}^{k} \frac{\mathbf{x_c^T} A^{-} \mathbf{x_c}}{\mathbf{x_c^T x_c}}$.*

Definition 3.12. *[4] The **positive ratio cut** minimizes the number of positive edges between clusters. It is given by $\min_{\{\mathbf{x_1},\ldots,\mathbf{x_k}\}\in I} \sum_{c=1}^{k} \frac{\mathbf{x_c^T} L^{+} \mathbf{x_c}}{\mathbf{x_c^T x_c}}$.*

Definition 3.13. *[4] The **negative ratio cut** maximizes the number of negative edges between clusters. It is given by $\max_{\{\mathbf{x_1},\ldots,\mathbf{x_k}\}\in I} \sum_{c=1}^{k} \frac{\mathbf{x_c^T} L^{-} \mathbf{x_c}}{\mathbf{x_c^T x_c}}$.*

Definition 3.14. *[4] The **balance ratio cut** combines the positive ratio cut with the negative ratio association and simultaneously minimizes the number of positive edges between clusters while minimizing the number of negative edges within clusters. It is given by* $\min_{\{\mathbf{x_1},\ldots,\mathbf{x_k}\}\in I} \sum_{c=1}^{k} \frac{\mathbf{x_c^T}(D^{+}-A)\mathbf{x_c}}{\mathbf{x_c^T}\mathbf{x_c}}$.

Definition 3.15. *[4] The **balance ratio association** combines the negative ratio cut with the positive ratio association and simultaneously maximizes the number of positive edges within clusters while maximizing the number of negative edges between clusters. It is given by* $\max_{\{\mathbf{x_1},\ldots,\mathbf{x_k}\}\in I} \sum_{c=1}^{k} \frac{\mathbf{x_c^T}(D^{-}+A)\mathbf{x_c}}{\mathbf{x_c^T}\mathbf{x_c}}$.

Definition 3.16. *[4] The **balance normalized cut** is very similar to the balance ratio cut, except it normalized the clusters by volume instead of number of nodes. It is given by* $\min_{\{\mathbf{x_1},\ldots,\mathbf{x_k}\}\in I} \sum_{c=1}^{k} \frac{\mathbf{x_c^T}(D^{+}-A)\mathbf{x_c}}{\mathbf{x_c^T}\bar{D}\mathbf{x_c}}$

Definition 3.17. *[4] Similarly, the **balance normalized association** can be derived from the balance ratio association and is given by* $\max_{\{\mathbf{x_1},\ldots,\mathbf{x_k}\}\in I} \sum_{c=1}^{k} \frac{\mathbf{x_c^T}(D^{-}+A)\mathbf{x_c}}{\mathbf{x_c^T}\bar{D}\mathbf{x_c}}$

Finally, Chiang et al prove the following:

Theorem 3.5. *[4] Minimizing balance normalized cut is equivalent to maximizing balance normalized association.*

Thus, the choice between balance normalized cut and association is inconsequential.

Zheng et al build on Kunegis et al, but they use an embedding map rather than an index of partitions. This yields additional information on similarity between nodes rather than simply assigning cluster labels. Additionally, the authors argue that an embedding map is more likely to yield an approximate global solution rather than local optima.

Zheng et al take a two-step approach: (1) the Rayleigh quotient of the random walk normalized Laplacian is used as an objective function to achieve the embedding, and (2) an objective function derived from the normalized signed cuts in [14] is used to complete clustering.

Mercado et al seek to address shortcomings in the previous adaptations of spectral clustering to signed graphs. Specifically, they speculate that prior technique's inability to recover ground-truth labels in real data sets is because they use a arithmetic mean of the positive-edge and negative-edge Laplacians [17], which introduces noise to the embedding of the data points. Furthermore, with the arithmetic mean the smallest eigenvectors of the Laplacian do not necessarily correspond to the smallest eigenvalues. The authors propose the use of the geometric mean of the positive-edge and negative-edge Laplacians to remedy these issues, although they concede that the geometric mean is much more computationally expensive than the arithmetic mean and not well-suited to large, sparse networks. Building on their 2016 paper, Mercado et al. further modified the Laplacian in their 2019 publication by combining the positive and negative Laplacians using the matrix power mean [18]. This approach further improved the results, as we will demonstrate in experiment 1.

In 2019, Cucuringu et al. introducing the SPONGE algorithm, which uses optimization to find the k smallest eigenvectors before applying signed graph clustering techniques [5].

4 Signed graph methodologies

Early Work

While modified methods borrows from unsigned graph clustering remaining among the most popular clustering tools to this day, they are not the only successful approach. Starting in the 1970s, researchers have developed tools specifically for signed graphs that cannot be applied directly to unsigned graphs. Early work in this field was heavily constrained by computational technology, and thus focused on smaller, denser networks. While effective at the time, these methodologies do not necessarily translate well to the large, sparse networks that are the focus of most modern research.

Breiger, et al. build on the concept of blockmodels to develop their system for clustering signed data [3]. To understand a blockmodel, we must first begin with an algebraic definition.

Definition 4.1. *Let S be a set and let $\{R_i\}_{i=1}^m$ be a set of binary relations on S. Individuals $a, b \in S$ are said to be **structurally equivalent** if and only if for any $c \in S$ and any $R_i \in \{R_i\}_{i=1}^m$, $aR_ic \iff bR_ic$ and $cR_ia \iff cR_ib$.*

In graph theory terms, we can understand structural equivalence of two nodes to mean that both nodes are adjacent to exactly the same set of nodes with the same edge weights. Since this occurrence is very rare in real data sets, the authors used the concept of a block model to relax the definition of structural equivalence.

The idea of a block model depends on rearranging the order of the nodes in the adjacency matrix. If the same permutation is applied to both the rows and the columns, the underlying network structure is not changes. Block models seek to permute the data in such a way that submatrices of all 0s exist within the adjacency matrix. The adjacency matrix is then divided into blocks, with a block being assigned a value of 0 if all edges within it are 0 are 1 otherwise. Nodes in the same block are assumed to be equivalent if the value of the block is equivalent, thus relaxing the prior definition of structural equivalence to fit real-world data sets.

Definition 4.2. *A blockmodel is said to be a **lean fit** to a matrix M if and only if there exists a permutation of M, yielding a permuted matrix M^* that can be blocked in such a way that: (1) Zeroblocks in M^* correspond to 0's in the blockmodel; and (2) Blocks in M^* containing at least one nonzero value have a corresponding value of 1 in the blockmodel.*

While a lean fit falls short of the algebraic definition of structural equivalence, the authors rationalized this decision by arguing that maintaining a social tie requires effort, while no work is required in the absence of a tie, thus it is appropriate to assign any block with nonzero values an overall value of 1 in the blockmodel. The authors further emphasize that nonzero blocks do not need to be true cliques, or fully connected subgraphs.

Several issues are immediately obvious with block models. First, for a graph with n vertices, there are $n!$ possible permutations (and thus blockmodels) of the vertices. Thus, exhaustively checking all possible blockmodels quickly becomes impractical even on relatively small graphs. Second, once a blockmodel is chosen the blockings must still be ennumerated and checked for lean fit. Third, care must be taken to ensure that any blocking has a reasonably small number of blocks and that the clustering is interpretable. Breiger et al. seek to resolve the third issue with their convergence of iterated correlations algorithm (CONCOR), which repeatedly applies bipartitions to the raw data until a hierarchical clustering at the appropriate level of granularity is established. CONCOR achieves this by repeatedly computing ordinary product moment correlation coefficients between the columns of the input matrix and storing them in a correlation matrix, then repeating the process on the correlation matrix until convergence is reached. Once convergence is reached, the bipartition is clear. The process can be repeated on each partition to establish further clusters.

CONCOR is unique in that it does not seek to optimize a specific metric, unlike many other clustering techniques.

Modularity-Based Methods

Modularity based metrics seek densely connected clusters using optimization techniques. In 2004, Newman et al. introduced a hierarchical agglomeration algorithm based on modularity for community detection [21]. It was notable for running in near linear time - a rare feat in graph theory - but did not apply to signed graphs. In 2007, Yang et al. introduced FEC, an algorithm for signed graph that was designed for densely connected networks [28]. FEC treats sign and density of edges as clustering attributes and can be used on both signed and unsigned graphs. In 2009, Gomez et al refined modularity metrics and extended existing methods to signed graphs that were directed, weighted, or contained loops [9].

Random Walk Models

Random walk models are a relatively new and unexplored area of signed graph clustering. The first random walk clustering algorithm was introduced by Harel, et al in 2001 for positively weighted graphs. An adaptation of the random walk methodology to signed graphs was not published until the Fast Clustering for Signed Graph (FCSG) algorithm was published by Hua, et al in 2020 [12].

Hua et al. contribute a novel approach to signed graph clustering using random walks. They define the random walk gap as the difference in cumulative transition probabilities between nodes in the positive-only subgraph versus the unsigned graph. The differences in transition probabilities are then used to adjust the positive edge weights of the original graph repeatedly.

5 Experimental Comparison

5.1 A Comparison of Ground-Truth Recovery for Modified Spectral Clustering Techniques

This experiment aims to identify the strengths and weaknesses of nine popular clustering techniques for signed graphs. We are specifically looking at graphs with known ground-truth community levels and assessing each algorithms ability to recover the true communities with respect to network characteristics including size, sparseness, balance, and the percent of edges that are positive.

Four relatively small graphs are presented: (1) Highland Tribes, which models agreeable and antagonistic relationships between tribes in the Eastern Central Highlands of New Guinea; (2) Sampson's Monastery, which models sentiment over time between novice monks in a New England monastery; (3) Football, which tracks twitter interactions between 248 players belonging to 20 distinct clubs of the English Premier League and has been augmented with negative edges; and (4) Olympics, which models Twitter interactions between 464 athletes belonging to 28 distinct sports involved in the 2012 London Olympics and has also been augmented with negative edges.

Nine different clustering methods for signed graphs are applied to each data set: spectral clustering using the signed graph Laplacian [14], spectral clustering using the symmetric Laplacian, spectral clustering using the symmetric separated Laplacian [14], balanced normalized cuts, symmetric balanced normalized cuts, SPONGE [5], symmetric SPONGE [5], geometric means [17], and matrix power means [18]. The Python package ”signet” was used to run the first 7 methodologies, while code provided by the authors of [17] and [18] was used for the final two algorithms.

The adjusted Rand index (ARI) was used to compare generated community labels with the ground-truth communities. The ARI ranges from -1 to 1, with a value near 0 representing a random pairing and 1 representing perfect recovery. Additionally, for each data set each method was compared against all others to detect the most (dis)similar methods.

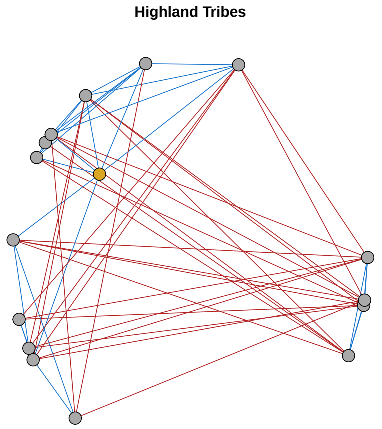
Highland Tribes and Sampson

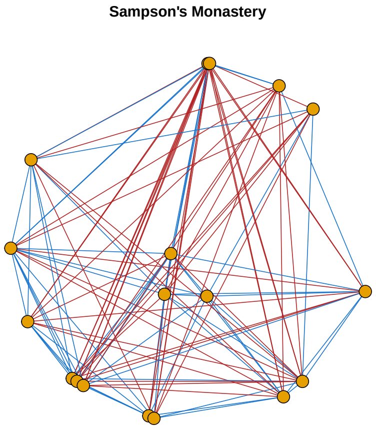
The Highland Tribes and Sampson data sets are both relatively small and dense signed graphs. The table below summarizes graph characteristics for each. Sparseness is calculates as the number of edges divided by the maximum possible number of edges excluding duplicate edges and loops. Balance is the percent of triangles in the graph that are balanced, and positive edge percent is the number of positive edges divided by the total number of edges.

	n nodes	n edges	sparseness	balance	pos edge %
highland	16	58	0.483	0.868	0.500
sampson	18	112	0.732	0.603	0.544

There are three communities present in Highland Tribes, with one node officially belonging to two clusters. The node in marked in gold in the figure below and, for the sake of the analysis, was labeled as belonging to the cluster with which it had the most positive ties. This node is also notable for being the only one in the data set to have no negative edges adjacent to it. As we can see from the figure below, Highland Tribes can almost be partitioned into a Harary cut and most of it's triangles are balanced. (Chen 2014) (Read 1954)

Sampson's monastery is comprised of eighteen monks divided into four non-overlapping groups, as identified by Samuel Franklin during a sociological study in 1968: the Young Turks, the Loyal Opposition, the Waverers, and the Outcasts (Franklin, 1968).


Highland Tribes

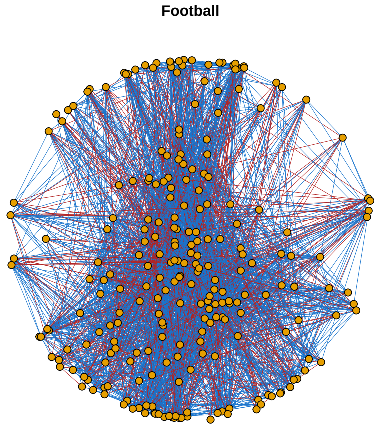

Sampson's Monastery

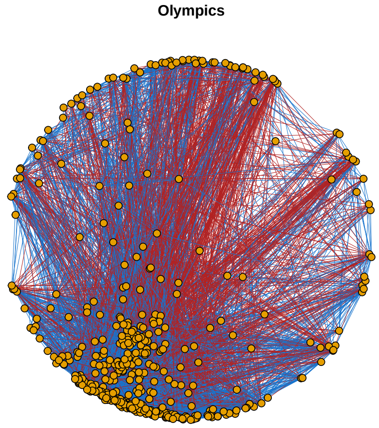
Football and Olympics

Negative edge augmentation was done using randomized block sampling, i.e. applying simple random selection to the communities to select a pair, and then applying simple random selection within that community to select nodes. A negative edge was then added between the two nodes. It should be noted that this approach is more likely to select nodes in smaller communities, and this was done intentionally due to the nature of the data sets. In the types of sports being modeled, competition is between clubs and teams rather than individual players. Thus, it is appropriate to use randomized block sampling as opposed to simple random sampling where all pairs of players from different clubs/sports are equally likely to be chosen. All generated random edges were checked against previously generated negative edges and existing positive edges to avoid duplication. The number of negative edges to add was determined as a percent relative to the number of positive edges - in this case, the number of negative edges was set to be 20 percent the number of positive edges. Negative edges are only added between communities, not within.

	n nodes	n edges	sparseness	balance	pos edge %
football	248	3174	0.104	0.878	0.833
olympics	464	9345	0.087	0.920	0.833

As we can see from the table above, these data sets are both larger and much sparser than Highland Tribes and Sampson. They also both have very high balance.

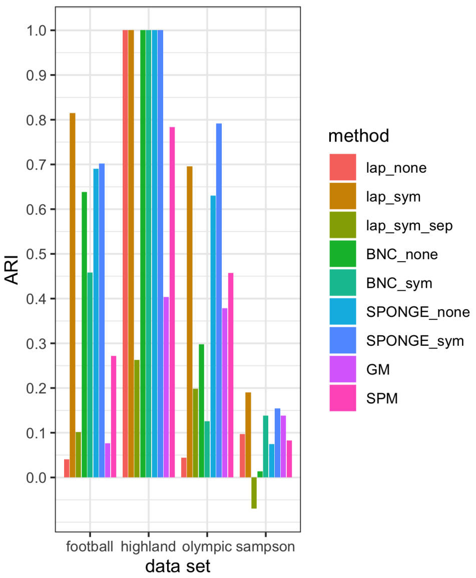

Football

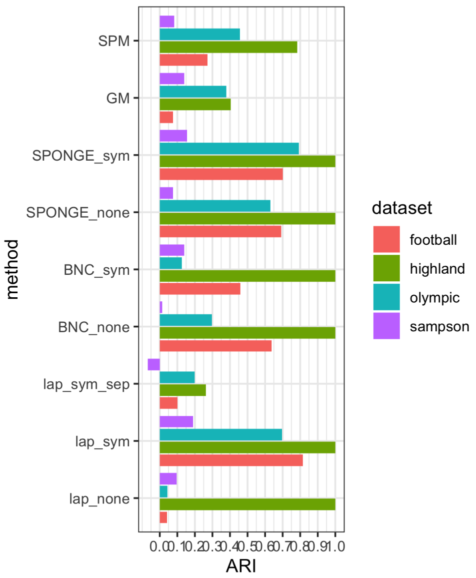

Olympics

Clustering

For each community, labels were generated for each of the nine clustering methods for k as given in the ground truth data set. After clustering, the adjusted Rand index (ARI) was computed for each labeling and used to compare the accuracy of the methods (Table 1). Additionally, for each data set each labeling was compared with all others to determine similarity in methods. The results of this analysis will be included in the final version of this manuscript.

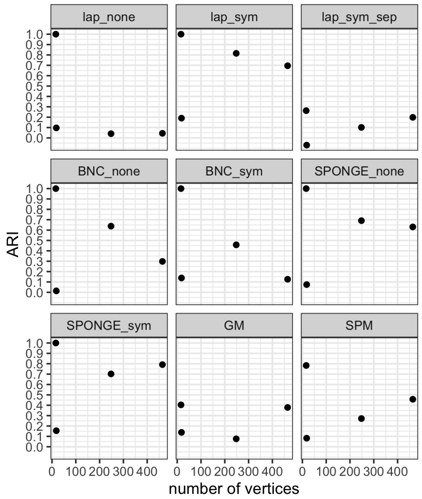
	lap	lap_sym	lap_sym_sep	BNC	BNC_sym	SPONGE	SPONGE_sym	GM	SPM
sampson	0.097	0.138	-0.045	0.156	0.093	0.075	0.109	0.138	0.083
olympics	0.016	0.727	0.168	0.357	0.105	0.600	0.822	0.379	0.457
football	0.049	0.781	0.087	0.603	0.472	0.691	0.703	0.076	0.272

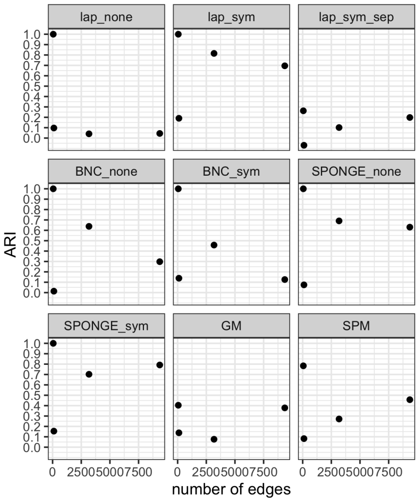




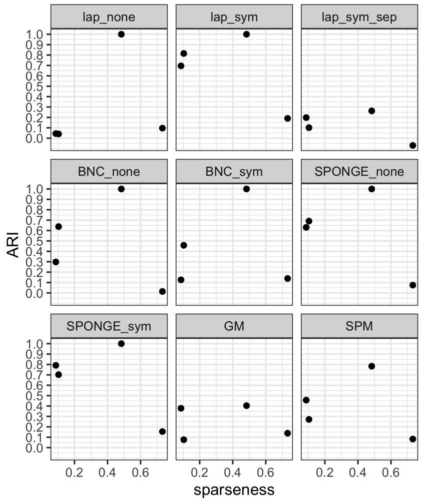
The first important set of results can be seen in the above plots. First, we notice that most algorithms were able to recover all or most of Highland Tribe's ground-truth community labels. This is likely due to the data sets small size and high balance. Sampson, on the other hand, performed poorly for all tested algorithms. Again, this is likely due to a combination of size and low balance in the set creating noise that the clustering techniques could not decipher. Strong performers across all data sets were the symmetric Laplacian, SPONGE, and symmetric SPONGE.

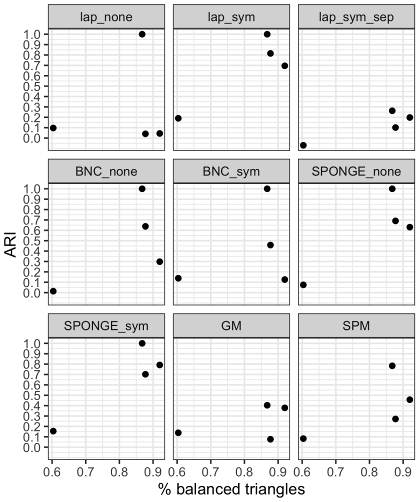
Next, we directly assessed clustering success with respect to the data sets' size, sparseness, balance, and percent of positive edges.



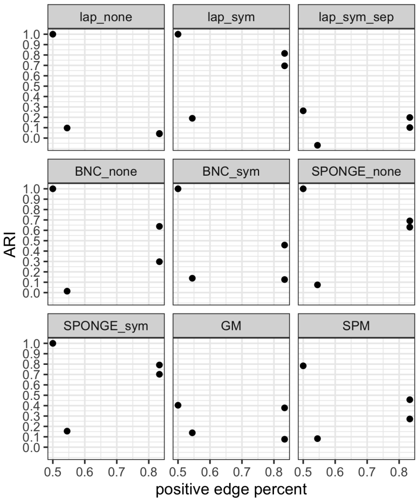


We used the number of vertices and the number of edges to measure the size of the graph. From the figures above, there are no clear trends in the effectiveness of a clustering algorithm with respect to size of the data set within the scope of this experiment.





Next, we assess clustering success with respect to sparseness and balance. While a clear pattern does not emerge for sparseness, we can see that the more balanced data sets tend to give better results across most methods, and especially for the three that we have previously identified as top performers (symmetric Laplacian, SPONGE, and symmetric SPONGE). It is also worth noting that SPM seems more reactive to balance than GM. SPM was an update to the methodology outlined in GM, suggesting the matrix power mean does a better job on weakly balanced graphs than the geometric mean.



The last graph characteristic we assessed was the percent of positive edges. Again, a clear pattern was not observed here and the top performing algorithms did not seem to be affected by the percent of positive edges. Any effect that was present was likely due to a confounding effect with balance.

5.2 A Comparison of Scalability for Benchmark Clustering Techniques

In experiment 2, we select leading algorithms from each of the major underlying methodologies: adapted spectral clustering (using the signed symmetric graph Laplacian), matrix power means, FEC, RWG, and SPONGE. We assess clustering on three real data sets: Wikipedia Elections [16], Slashdot [16], and Correlates of War [25]. For each data set, the number of clusters is chosen using the eigengap method [26] and each algorithm is applied using the same value of k. We compare the clustering time for each data set across algorithm assess clustering success with blockmodel diagrams and the Q modularity metric proposed in (Gomez 2009).

Definition 5.1. *The modularity metric Q is given by*
$Q = \frac{m^+}{m^+ + m^-}Q^+ - \frac{m^-}{m^+ + m^-}Q^-$*, where*
$Q^+ = \frac{1}{m^+}\sum_i \sum_j (w_{ij}^+ - \frac{w_i^+ w_j^+}{m^+})\delta(c_i, c_j)$*, and*
$Q^- = \frac{1}{m^-}\sum_i \sum_j (w_{ij}^- - \frac{w_i^- w_j^-}{m^-})\delta(c_i, c_j)$

The authors of FEC, SPONGE, matrix power means provided code with their publications which was used for these experiments. Spectral clustering with the signed Laplacian was done using the 'signet' package in Python. Since no code was provided for the FCSG random walk gap method, we implemented it in Python.

During the implementation of the FCSG algorithm, several issues that were not addressed in the paper were discovered. Once significant limitation of the random walk gap algorithm that was uncovered during this analysis was the assumption of the "small world hypothesis", i.e. the theory that in a social network, most users are linked by no more than 5 degrees of separation. The parameter used in the random walk gap matrix calculation, L, must be greater than or equal to the diameter of the positive-edge-only subgraph of G. The authors recommend that L be set to 5, and warn that the algorithm begins to degrade in quality if L is greater than or equal to 10. For highly sparse graphs, the diameter can exceed 10 and thus the recommended value of L is mathematically impossible to use. Additionally, we found potential to speed up the algorithm during the greedy clustering subroutine by allowing positive edges to remain in the final graph if they are below a specified cutoff.

While the issues described above with FCSG were resolved and lead to improvements in the algorithm, they have not been resolved in a scalable manner. Since the original FCSG code was not published were were unable to verify the run-times in the original paper and were not able to run it on any of the experiment 2 data sets without the use the a high performance computing cluster. We did manage to get FCSG to run on Highland Tribes as a smaller proof of concept data set. Without cluster labels from FCSG for Wikipedia, Slashdot, and Correlates of War the scalability branch of this study could not be completed, and has been delayed until FCSG can be run on a high performance computing cluster.

6 Future Work

The next step in this project will be to run FCSG on a high performance computing cluster (LEAP). Once we have labels with run-times for the large sample data-sets, we will compare spectral clustering with the signed symmetric Laplacian, matrix power means, FEC, FCSG, and SPONGE using blockmodels and the Q modularity metrics. Ultimately, the purpose of this study has been to establish the current state of the field with the goal of assessing our own suite of signed graph data mining techniques that are based on balance cut metrics. We have had early success using traditional spectral clustering on a feature vector of metrics generated by the balanced cut procedure described in [23] that we intend to fully develop into a community detection technique in its own right, and will integrate it into both experiment 1 and experiment 2 as another leading technique for signed graph spectral clustering.

References

[1] T. Antal, P. L. Krapivsky, and S. Redner. “Social Balance on Networks: The Dynamics of Friendship and Enmity”. In: *Physica D: Nonlinear Phenomena* 224.1 (Dec. 2006), pp. 130–136. ISSN: 01672789. DOI: 10.1016/j.physd.2006.09.028. arXiv: physics/0605183. URL: http://arxiv.org/abs/physics/0605183 (visited on 04/16/2021).

[2] Robert Axelrod and D. Scott Bennett. “A Landscape Theory of Aggregation”. In: *British Journal of Political Science* 23.2 (1993), pp. 211–233. DOI: 10.1017/S000712340000973X.

[3] Ronald L Breiger, Scott A Boorman, and Phipps Arabie. “An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling”. In: *Journal of Mathematical Psychology* 12.3 (1975), pp. 328–383. ISSN: 0022-2496. DOI: https://doi.org/10.1016/0022-2496(75)90028-0. URL: https://www.sciencedirect.com/science/article/pii/0022249675900280.

[4] Kai-Yang Chiang, Joyce Jiyoung Whang, and Inderjit S. Dhillon. “Scalable clustering of signed networks using balance normalized cut”. In: *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM ’12*. Maui, Hawaii, USA: ACM Press, 2012, p. 615. ISBN: 978-1-4503-1156-4. DOI: 10.1145/2396761.2396841. URL: http://dl.acm.org/citation.cfm?doid=2396761.2396841 (visited on 08/18/2020).

[5] Mihai Cucuringu et al. “SPONGE: A generalized eigenproblem for clustering signed networks”. In: *arXiv:1904.08575 [cs, math, stat]* (May 19, 2019). arXiv: 1904.08575. URL: http://arxiv.org/abs/1904.08575 (visited on 09/16/2020).

[6] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. “Kernel k-means: spectral clustering and normalized cuts”. In: *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’04*. the 2004 ACM SIGKDD international conference. Seattle, WA, USA: ACM Press, 2004, p. 551. DOI: 10.1145/1014052.1014118. URL: http://portal.acm.org/citation.cfm?doid=1014052.1014118 (visited on 04/12/2021).

[7] Pouya Esmailian and Mahdi Jalili. “Community Detection in Signed Networks: the Role of Negative ties in Different Scales”. In: *Scientific Reports* 5.1 (Nov. 2015), p. 14339. ISSN: 2045-2322. DOI: 10.1038/srep14339. URL: http://www.nature.com/articles/srep14339 (visited on 08/27/2020).

[8] M. Girvan and M. E. J. Newman. “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences* 99.12 (June 11, 2002), pp. 7821–7826. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.122653799. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.122653799 (visited on 04/16/2021).

[9] Sergio Gómez, Pablo Jensen, and Alex Arenas. “Analysis of community structure in networks of correlated data”. In: *Phys. Rev. E* 80 (1 July 2009), p. 016114. DOI: 10.1103/PhysRevE.80.016114. URL: https://link.aps.org/doi/10.1103/PhysRevE.80.016114.

[10] F. Harary. "On the notion of balance of a signed graph". In: *Michigan Math. J.* 2(2) (1953), pp. 143–146.

[11] F. Heider. "Attitudes and cognitive organization". In: *J. Psychology* 21 (1946), pp. 107–112.

[12] Jialin Hua, Jian Yu, and Miin-Shen Yang. "Fast clustering for signed graphs based on random walk gap". en. In: *Social Networks* 60 (Jan. 2020), pp. 113–128. ISSN: 03788733. DOI: 10.1016/j.socnet.2018.08.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378873317302460 (visited on 08/27/2020).

[13] Arzum Karatas and Serap Sahin. "Application Areas of Community Detection: A Review". In: *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*. 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT). ANKARA, Turkey: IEEE, Dec. 2018, pp. 65–70. ISBN: 978-1-72810-472-0. DOI: 10.1109/IBIGDELFT.2018.8625349. URL: https://ieeexplore.ieee.org/document/8625349/ (visited on 04/16/2021).

[14] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. "The Slashdot Zoo: Mining a Social Network with Negative Edges". In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. Madrid, Spain: ACM, 2009.

[15] Jérôme Kunegis et al. "Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization". en. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Apr. 2010, pp. 559–570. ISBN: 978-0-89871-703-7 978-1-61197-280-1. DOI: 10.1137/1.9781611972801.49. URL: https://epubs.siam.org/doi/10.1137/1.9781611972801.49 (visited on 08/18/2020).

[16] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data. June 2014.

[17] Pedro Mercado, Francesco Tudisco, and Matthias Hein. "Clustering Signed Networks with the Geometric Mean of Laplacians". In: (), p. 9.

[18] Pedro Mercado, Francesco Tudisco, and Matthias Hein. "Spectral Clustering of Signed Graphs via Matrix Power Means". In: (), p. 28.

[19] Michael Moore. "An international application of Heider's balance theory". In: *European Journal of Social Psychology* 8.3 (1978), pp. 401–405. DOI: https://doi.org/10.1002/ejsp.2420080313. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420080313. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2420080313.

[20] Michael Moore. "Structural balance and international relations". In: *European Journal of Social Psychology* 9.3 (1979), pp. 323–326. DOI: https://doi.org/10.1002/ejsp.2420090309. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420090309. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2420090309.

[21] M. E. J. Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8577–8582. ISSN: 0027-8424. DOI: 10.1073/pnas.0601602103. eprint: https://www.pnas.org/content/103/23/8577.full.pdf. URL: https://www.pnas.org/content/103/23/8577.

[22] Kenneth Read. "Cultures of the Central Highlands, New Guinea". In: *Southwestern Journal of Anthropology* 10.1 (1954), pp. 1–43.

[23] Lucas Rusnak and Jelena Tešić. *Characterizing Attitudinal Network Graphs through Frustration Cloud*. 2021. arXiv: 2009.07776 [cs.SI].

[24] Majid Saberi et al. "Topological impact of negative links on the stability of resting-state brain network". In: *Scientific Reports* 11.1 (Dec. 2021), p. 2176. ISSN: 2045-2322. DOI: 10.1038/s41598-021-81767-7. URL: http://www.nature.com/articles/s41598-021-81767-7 (visited on 04/16/2021).

[25] Meredith Reid Sarkees and Frank Wayman. "Resort to War: 1816 - 2007." In: *Washington DC: CQ Press.* (2010).

[26] Jianbo Shi and Jitendra Malik. "Normalized Cuts and Image Segmentation". In: *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 22.8 (2000), p. 18.

[27] Jiliang Tang et al. “A Survey of Signed Network Mining in Social Media”. en. In: *arXiv:1511.07569 [physics]* (June 2016). arXiv: 1511.07569. URL: `http://arxiv.org/abs/1511.07569` (visited on 08/27/2020).

[28] Bo Yang, William Cheung, and Jiming Liu. “Community Mining from Signed Social Networks”. In: *IEEE Transactions on Knowledge and Data Engineering* 19.10 (2007), pp. 1333–1348. DOI: `10.1109/TKDE.2007.1061`.