Scalable Clustering of Signed Networks Using Balance Normalized Cut

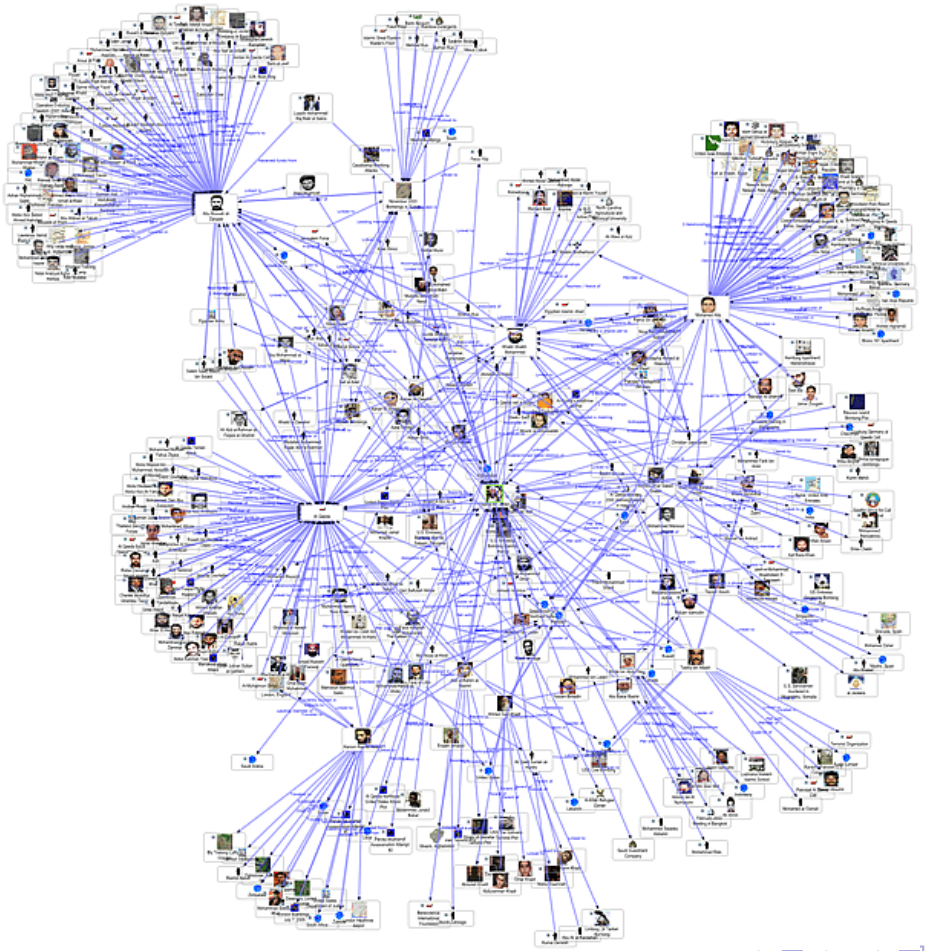
Kai-Yang Chiang, Joyce Jiyoung Whang, Inderjit S. Dhillon
University of Texas at Austin

The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)
Oct. 29 - Nov. 2, 2012

Contents

Introduction

- Social Networks
  - Nodes: the individual actors
  - Edges: the relationships (social interactions) between the actors



Introduction

- Signed Networks
  - Positive relationship: friendship, collaboration
  - Negative relationship: distrust, disagreement
- Clustering problem in signed networks
  - Entities within the same cluster have a positive relationship.
  - Entities between different clusters have a negative relationship.

- Contributions
  - New k-way objectives and kernels for signed networks.
  - Show equivalence between our new k-way objectives and a general weighted kernel k-means objective.
  - Fast and scalable clustering algorithm for signed networks.

Clustering of Unsigned Networks

Graph Cuts on Unsigned Networks

- **Ratio Cut objective**
  - Minimizes the number of edges between different clusters relative to the size of the cluster.
  - The graph Laplacian $L = D - A$ where $D_{ii} = \sum_{j=1}^{n} A_{ij}$.

$$\min_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T L \mathbf{x}_c}{\mathbf{x}_c^T \mathbf{x}_c} \right).$$

  - Under the special case $k = 2$,

$$\min_{\mathbf{x}} \left(\mathbf{x}^T L \mathbf{x} \right), \text{ where } x_i = \begin{cases} \sqrt{|\pi_2|/|\pi_1|}, & \text{if node } i \in \pi_1, \\ -\sqrt{|\pi_1|/|\pi_2|}, & \text{if node } i \in \pi_2. \end{cases}$$

Graph Cuts on Unsigned Networks

- **Ratio Association objective**
  - Maximizes the number of edges within clusters relative to the size of the cluster.

$$\max_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T A \mathbf{x}_c}{\mathbf{x}_c^T \mathbf{x}_c} \right), \text{ where } x_c(i) = \begin{cases} 1, & \text{if node } i \in \pi_c, \\ 0, & \text{otherwise.} \end{cases}$$

- **Normalized Association** and **Normalized Cut objectives**
  - Normalized by the volume of each cluster.
  - The volume of a cluster: the sum of degrees of nodes in the cluster.

$$\max_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T A \mathbf{x}_c}{\mathbf{x}_c^T D \mathbf{x}_c} \right) \equiv \min_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T L \mathbf{x}_c}{\mathbf{x}_c^T D \mathbf{x}_c} \right).$$

Weighted Kernel K-means

- **A general weighted kernel k-means objective** is equivalent to **a weighted graph clustering objective**. (Dhillon et al. 2007)
- Weighted kernel k-means
  - Objective

$$\min_{\pi_1 \ldots \pi_k} \sum_{c=1}^{k} \sum_{\mathbf{v}_i \in \pi_c} w_i \|\varphi(\mathbf{v}_i) - \mathbf{m}_c\|^2, \text{ where } \mathbf{m}_c = \frac{\sum_{\mathbf{v}_i \in \pi_c} w_i \varphi(\mathbf{v}_i)}{\sum_{\mathbf{v}_i \in \pi_c} w_i}.$$

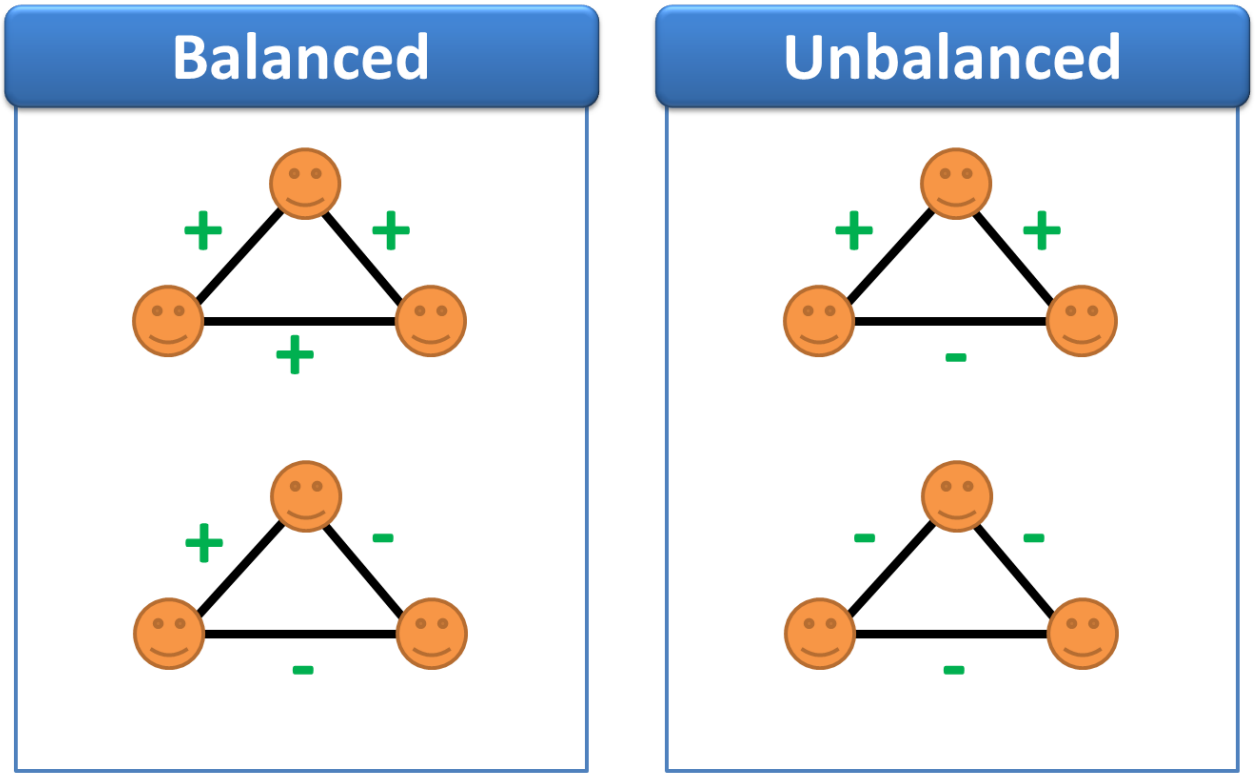
  - Algorithm
    - Computes the closest centroid for every node, and assigns the node to the closest cluster.
    - After all the nodes are considered, the centroids are updated.
    - Given the Kernel matrix K, where $K_{ji} = \langle \varphi(\mathbf{v}_j), \varphi(\mathbf{v}_i) \rangle$,

$$D(\mathbf{v}_i, \mathbf{m}_c) = K_{ii} - \frac{2 \sum_{j \in c} w_j K_{ji}}{\sum_{j \in c} w_j} + \frac{\sum_{j \in c} \sum_{l \in c} w_j w_l K_{jl}}{(\sum_{j \in c} w_j)^2}.$$

Signed Networks and Social Balance

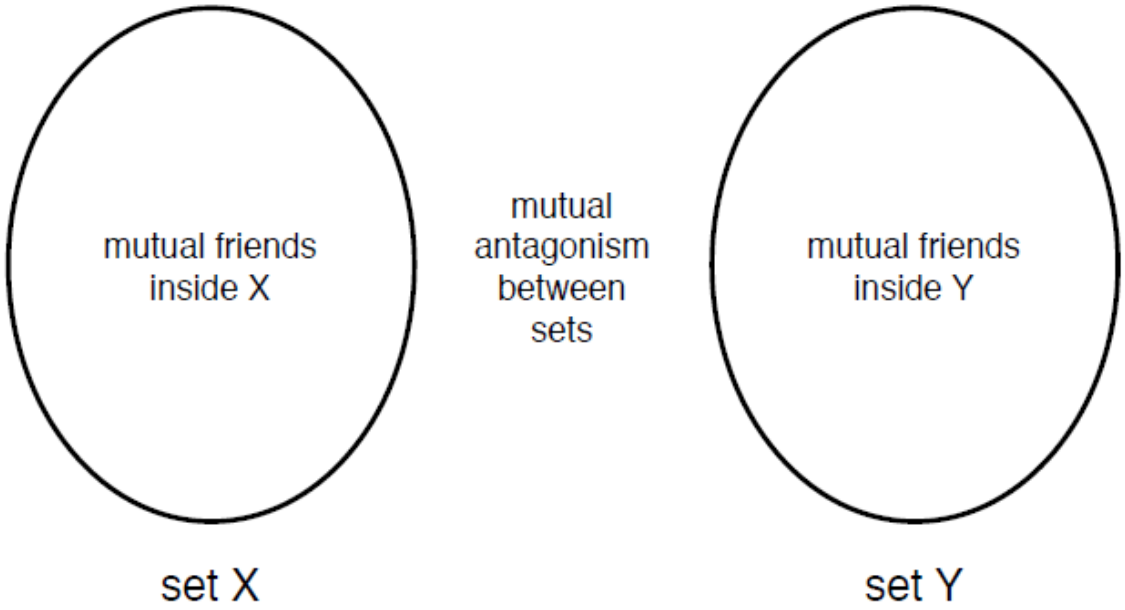
Social Balance

- Certain configuration of positive and negative edges are more plausible than others.
  - A friend of my friend is my friend.
  - An enemy of my friend is my enemy.
  - An enemy of my enemy is my friend.



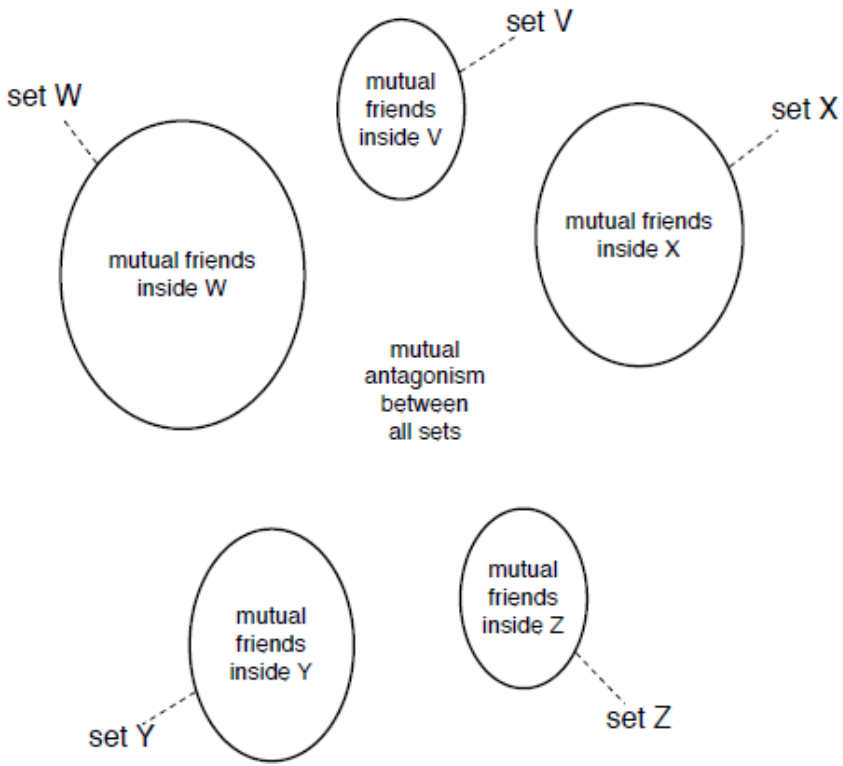
Balance Theory

- A network is balanced iff (i) all of its edges are positive, or (ii) nodes can be clustered into two groups such that edges within groups are positive and edges between groups are negative. (Cartwright and Harary)



Weak Balance Theory

- Allows an enemy of one's enemy to still be an enemy.
- A network is weakly balanced iff (i) all of its edges are positive, or (ii) nodes can be clustered into k groups such that edges within groups are positive and edges between groups are negative. (Davis 1967)



Clustering via Signed Laplacian

Signed Laplacian

- The signed Laplacian $\bar{L} = \bar{D} - A$
  where $\bar{D}$ is the diagonal absolute degree matrix, i.e., $\bar{D}_{ii} = \sum_{j=1}^{n} |A_{ij}|$.
  (Kunegis et al. 2010)
- $\bar{L}$ is always positive semidefinite: $\forall \mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^T \bar{L} \mathbf{x} = \sum_{(i,j)} |A_{ij}|(x_i - \text{sgn}(A_{ij})x_j)^2 \geq 0.$$

- ***k*-way ratio cut for signed networks**
  - The sum of positive edge weights for edges that lie between different clusters and the sum of negative edge weights of all edges lie within the same cluster, normalized by each cluster's size.

Signed Laplacian

- The 2-way signed ratio cut objective can be formulated as an optimization problem with a quadratic form:

$$\min_{\mathbf{x}} \left(\mathbf{x}^T \bar{L} \mathbf{x} \right),$$

where the 2-class indicator $\mathbf{x}$ has the following form:

$$x_i = \begin{cases} \frac{1}{2}(\sqrt{|\pi_2|/|\pi_1|} + \sqrt{|\pi_1|/|\pi_2|}), & \text{if node } i \in \pi_1, \\ -\frac{1}{2}(\sqrt{|\pi_2|/|\pi_1|} + \sqrt{|\pi_1|/|\pi_2|}), & \text{if node } i \in \pi_2. \end{cases}$$

Extension of Signed Laplacian to k-way Clustering

- Extension to k-way objective

$$\min_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T \bar{L} \mathbf{x}_c}{\mathbf{x}_c^T \mathbf{x}_c} \right).$$

Theorem

There does not exist any representation of $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ such that the objective minimizes the general k-way signed ratio cut.

- This direct extension suffers a weakness.
- No matter how we select an indicator vector, we will always punish some desirable clustering patterns.

k-way Signed Objectives for Clustering

Proposed k-way Signed Objectives

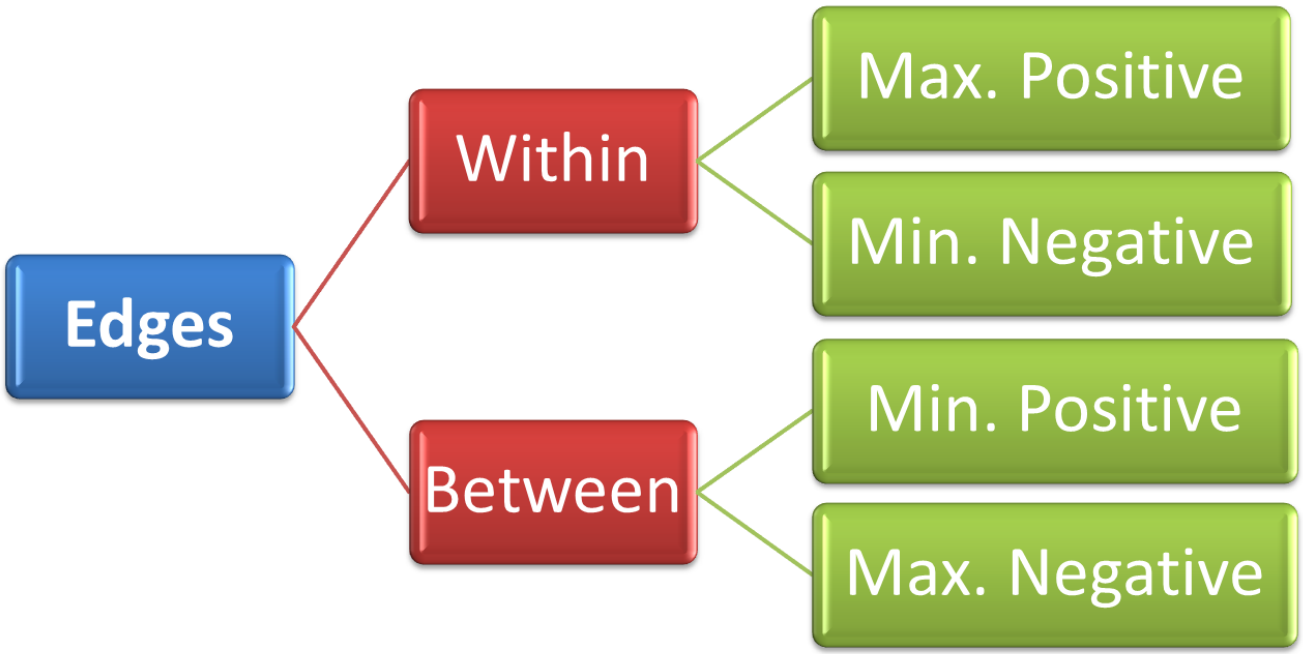
- Adjacency matrix of a signed network

$$A_{ij} \begin{cases} > 0, & \text{if relationship of } (i,j) \text{ is positive,} \\ < 0, & \text{if relationship of } (i,j) \text{ is negative,} \\ = 0, & \text{if relationship of } (i,j) \text{ is unknown.} \end{cases}$$

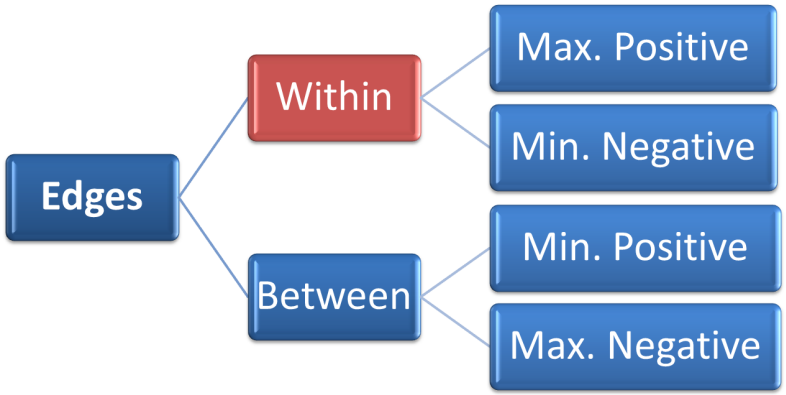
  - We can break A into its positive part A^+ and negative part A^-.
  - Formally, $A^+_{ij} = \max(A_{ij}, 0)$ and $A^-_{ij} = -\min(A_{ij}, 0)$.
  - By this definition, we have $A = A^+ - A^-$.

Proposed k-way Signed Objectives

- Overview of k-way signed objectives



Proposed k-way Signed Objectives



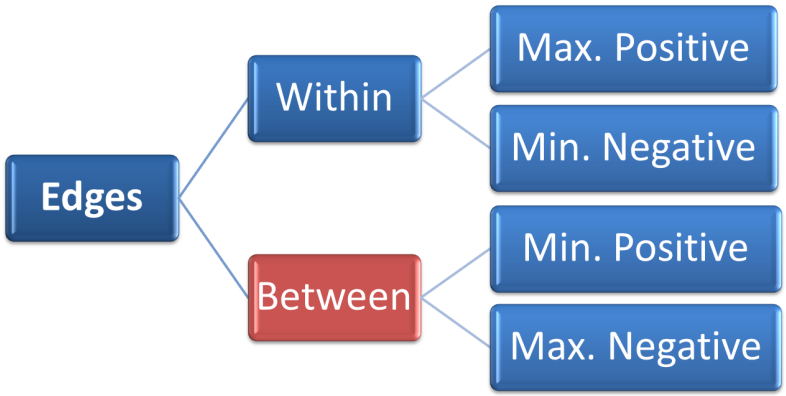
- **Positive/Negative Ratio Association**
  - Positive Ratio Association

$$\max_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T A^+ \mathbf{x}_c}{\mathbf{x}_c^T \mathbf{x}_c} \right).$$

  - Negative Ratio Association

$$\min_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T A^- \mathbf{x}_c}{\mathbf{x}_c^T \mathbf{x}_c} \right).$$

Proposed k-way Signed Objectives



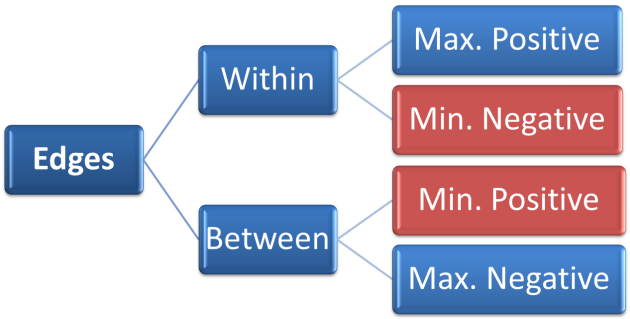
- **Positive/Negative Ratio Cut**
  - Positive Ratio Cut
    - Minimizes the number of positive edges between clusters.

$$\min_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I}\left\{\sum_{c=1}^{k}\frac{\mathbf{x}_c^T(D^+-A^+)\mathbf{x}_c}{\mathbf{x}_c^T\mathbf{x}_c}=\sum_{c=1}^{k}\frac{\mathbf{x}_c^T L^+\mathbf{x}_c}{\mathbf{x}_c^T\mathbf{x}_c}\right\},$$

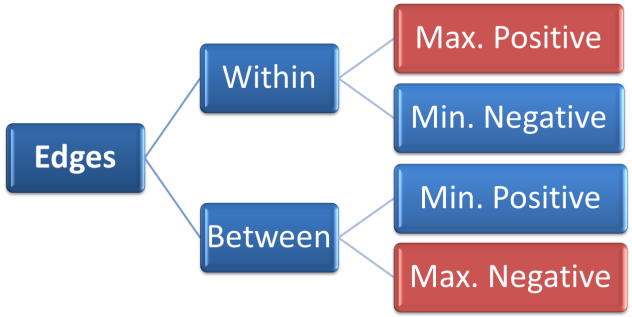
where D^+ is the diagonal degree matrix of A^+.

  - The Negative Ratio Cut can also be defined similarly.

Proposed k-way Signed Objectives



(a) Balance Ratio Cut



(b) Balance Ratio Association

- **Balance Ratio Cut/Association**
  - Balance Ratio Cut

$$\min_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T (D^+ - A)\mathbf{x}_c}{\mathbf{x}_c^T \mathbf{x}_c} \right).$$

  - Balance Ratio Association

$$\max_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T (D^- + A)\mathbf{x}_c}{\mathbf{x}_c^T \mathbf{x}_c} \right).$$

Proposed k-way Signed Objectives

- **Balance Normalized Cut**
  - Objectives normalized by cluster volume instead of by the number of nodes in the clusters.
  - Balance Normalized Cut

$$\min_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}\in I} \left(\sum_{c=1}^{k} \frac{\mathbf{x}_c^T (D^+ - A)\mathbf{x}_c}{\mathbf{x}_c^T \bar{D}\mathbf{x}_c} \right).$$

Theorem

Minimizing balance normalized cut is equivalent to maximizing balance normalized association.

Multilevel Approach for Large-scale Signed Graph Clustering

Equivalence of Objectives

- Equivalence between k-ways signed objectives and weighted kernel k-means objective

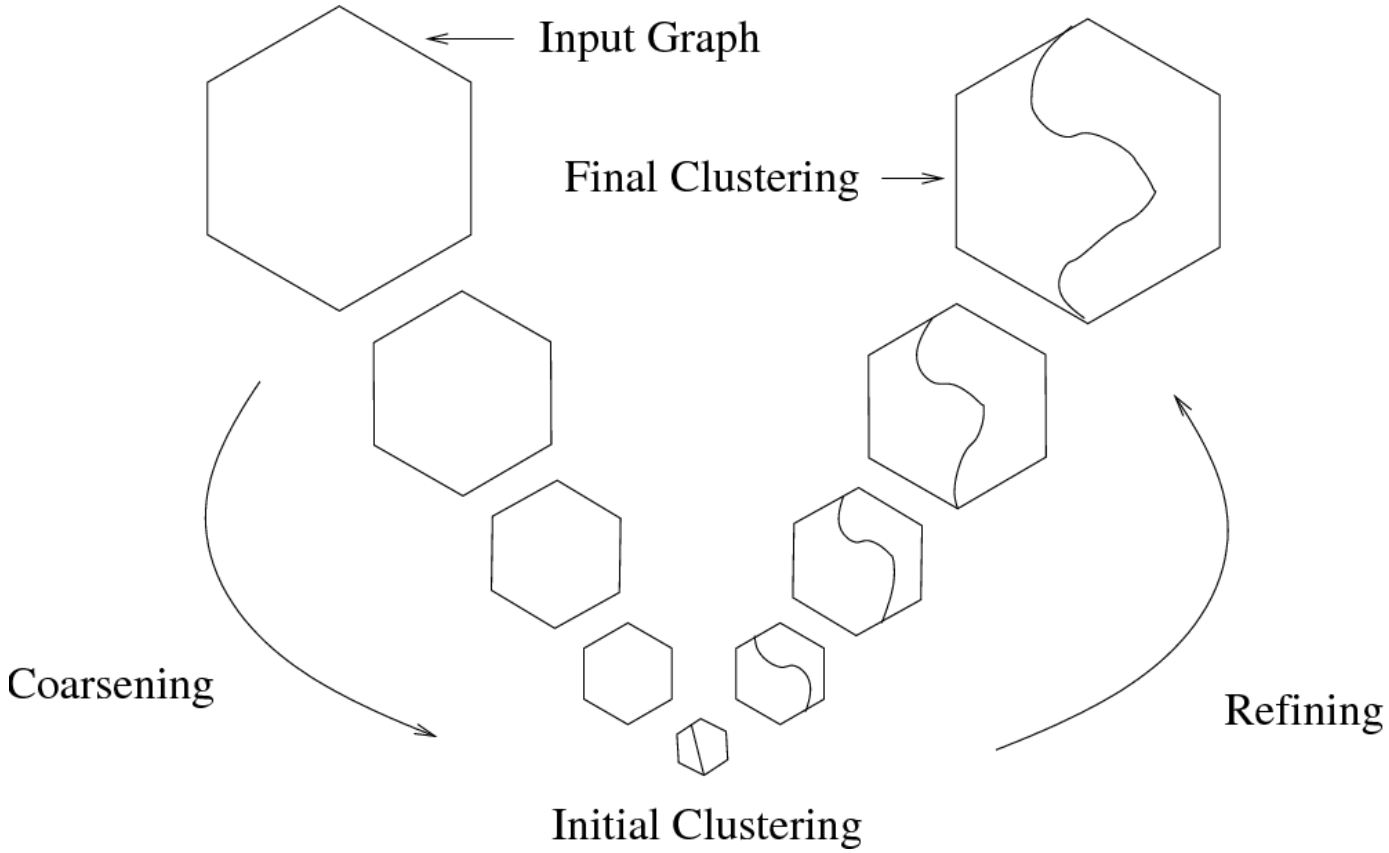
Theorem (Equivalence of objectives)

For any signed cut or association objective, there exists some corresponding weighted kernel k-means objective (with properly chosen kernel matrix), such that these two objectives are mathematically equivalent.

- We can use k-means like algorithm to optimize the objectives.
- Fast and scalable multilevel clustering algorithm for signed networks.

Multilevel Framework of Graph Clustering

- Overview



Multilevel Clustering Algorithm for Signed Networks

- Coarsening Phase
  - Given the input graph G_0, we generate a series of graphs $G_1 \ldots G_\ell$, such that $|V_{i+1}| < |V_i|$ for all $0 \leq i < \ell$.
- Base Clustering Phase
  - Minimize balance normalized cut of A^ℓ with spectral relaxation.
  - Perform unsigned graph clustering on A^{ℓ^+} using region-growing algorithm as in Metis. (Karypis and Kumar 1999)
- Refinement Phase
  - Derive clustering results in $G_{\ell-1}$, $G_{\ell-2}$, $\ldots$, G_0.
  - Given a clustering of G_i, the goal is to get a clustering result in G_{i-1}.
    - Project the clustering result in G_i to G_{i-1} as the initial clusters.
    - Refine the clustering result by running weighted kernel k-means.

Experimental Results

Graph Kernels

- Criteria and Kernels

Criterion	Kernel
Signed Laplacian	$\sigma I - \bar{L}$
Normalized Signed Laplacian	$\sigma \bar{D}^{-1} + \bar{D}^{-1} A \bar{D}^{-1}$
Positive Ratio Association	$\sigma I + A^{+}$
Positive Ratio Cut	$\sigma I - L^{+}$
Ratio Association	$\sigma I + A$
Balance Ratio Cut	$\sigma I - (D^{+} - A)$
Balance Normalized Cut	$\sigma \bar{D}^{-1} - \bar{D}^{-1}(D^{+} - A)\bar{D}^{-1}$

Table: Criteria and kernels considered in experiments.

Graph Kernels

- Experimental Setup and Metrics
  - Synthetic networks
    - Begin with a complete 5-weakly balanced network A_{com}, in which group sizes are $100, 200, \ldots 500$ respectively.
    - Uniformly sample some entries from A_{com} to form a weakly balanced network A, with two parameters: sparsity s and noise level ϵ.
  - Error rate
    - Have the "real" clustering as the ground truth in synthetic dataset

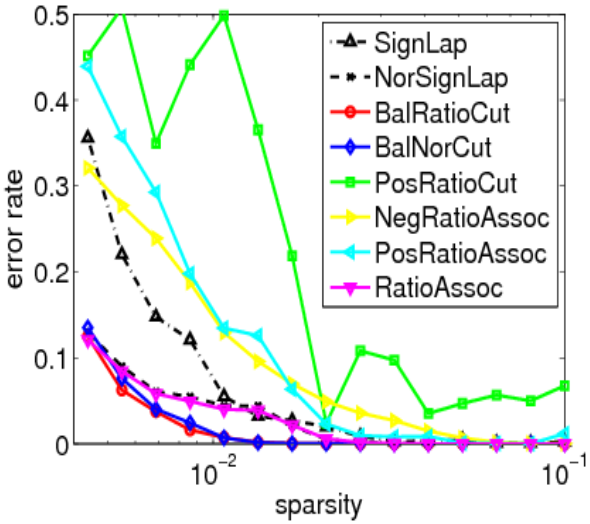
$$\sum_{c=1}^{k} \frac{\mathbf{x}_c^T A_{com}^- \mathbf{x}_c + \mathbf{x}_c^T L_{com}^+ \mathbf{x}_c}{n^2}.$$

  - Measuring the degree of imbalance of clusters
    - Ratio objective ($W = I$) and normalized objective ($W = \bar{D}$)

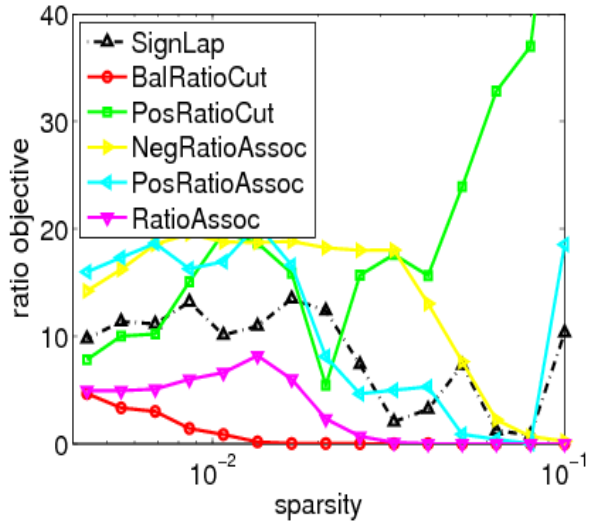
$$\sum_{c=1}^{k} \frac{\mathbf{x}_c^T A^- \mathbf{x}_c + \mathbf{x}_c^T L^+ \mathbf{x}_c}{\mathbf{x}_c^T W \mathbf{x}_c}.$$

Graph Kernels

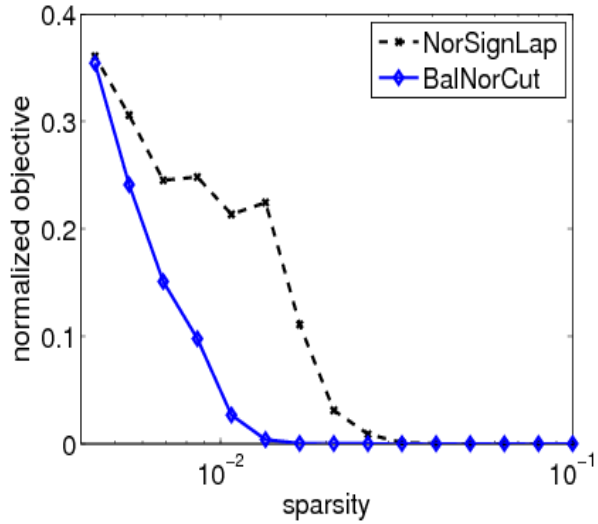
- Spectral clustering results using different kernels on weakly balanced networks, with different sparsity
  - PosRatioAssoc, NegRatioAssoc and PosRatioCut, which only consider one of positive or negative criterion, perform worse than others.
  - BalRatioCut and BalNorCut outperform SignLap and NorSignLap under every sparsity level.



(c) Error rate



(d) Ratio objective



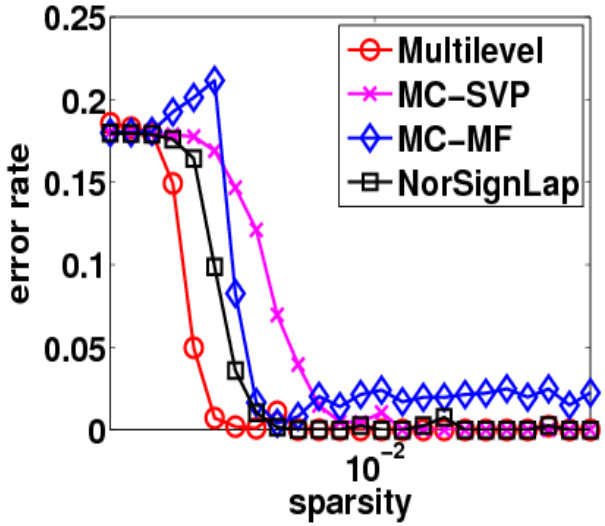
(e) Normalized objective

Multilevel Clustering

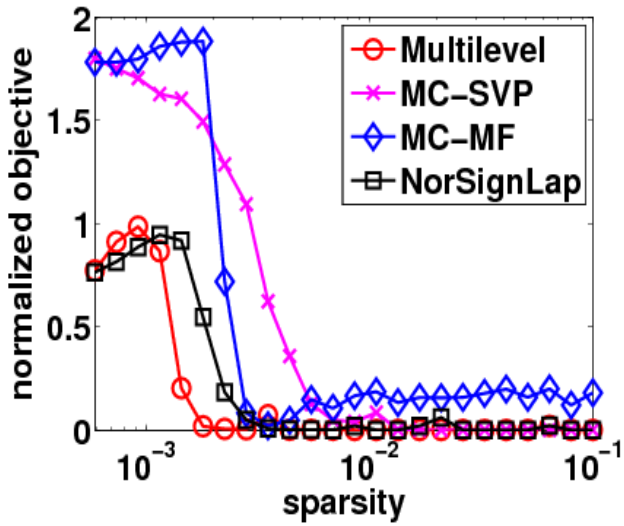
- Methods
  - Multilevel Clustering with Balance Normalized Cut
  - Normalized signed Laplacian (**NorSignLap**)
  - **MC-SVP**
    - Use SVP to complete the network and run k-means on k eigenvectors of the completed matrix to get the clustering result.
  - **MC-MF**
    - Complete the network using matrix factorization and derive two low rank factors $U, H \in \mathbb{R}^{n \times k}$, and run k-means on both U and H.
    - Select the clustering that gives us smaller normalized balance cut objective.

Multilevel Clustering

- Clustering results of multilevel clustering and other state-of-the-art methods on weakly balanced networks, with different sparsity
  - Create some networks sampled from a complete 10-weakly balanced network, in which each group contains $1,000$ nodes.
  - The multilevel clustering outperforms other state-of-the-art methods in most cases.



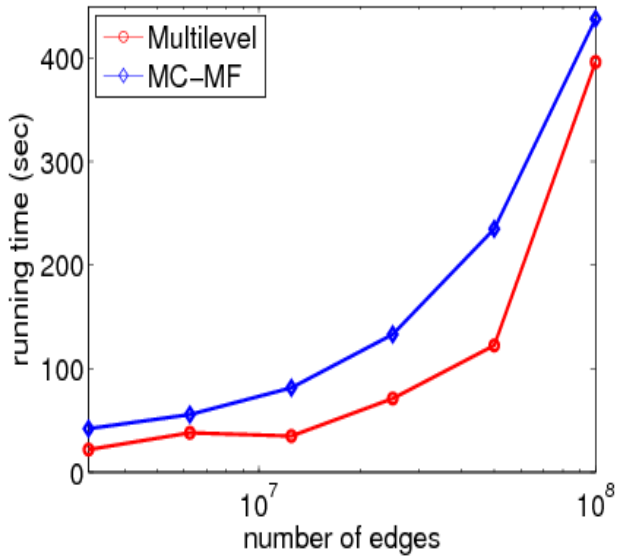
(f) Error rate



(g) Normalized objective

Multilevel Clustering

- Running time of multilevel clustering and MC-MF on weakly balanced networks
  - Consider A_{com} to be a large balanced network, which contains 20 groups, with $50,000$ nodes in each group.
  - Randomly sample some edges from A_{com} to form A with desired number of edges.
  - While we report the running time of whole procedure for multilevel clustering, we only report the time for computing two factors U and H for MC-MF.



Conclusions

- Show a fundamental weakness of the signed graph Laplacian in k-way clustering problems.
- New k-way objectives and kernels for signed networks.
- Equivalence between our new k-way objectives and a general weighted kernel k-means objective.
- Fast and scalable multilevel clustering algorithm for signed networks.
  - Comparable in accuracy to other state-of-the-art methods.
  - Much faster and more scalable.

References

- D. Cartwright and F. Harary. Structure balance: A generalization of Heiders theory. *Psychological Review*, 63(5):277293, 1956.
- J. A. Davis. Clustering and structural balance in graphs. *Human Relations*, 20(2):181187, 1967.
- I. S. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(11):1944-1957, 2007.
- F. Harary. On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2(2):143146, 1953.
- J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, E. W. D. Luca, and S. Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *SDM*, pages 559570, 2010.
- G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359392, 1999.
- J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, pages 641650, 2010.