Cluster Boosting and Data Discovery in Social Networks*

Maria Tomasso
Texas State University
San Marcos, Texas, USA
met48@txstate.edu

Lucas J. Rusnak
Texas State University
San Marcos, Texas, USA
lucas.rusnak@txstate.edu

Jelena Tešić
Texas State University
San Marcos, Texas, USA
jtesic@txstate.edu

ABSTRACT

We introduce a new ground-truth recovery boosting approach on antagonistic networks using the status-influence space obtained by the frustration cloud — a generalization of the frustration index which models nearest consensus-based states of a signed graph. The status and influence metrics are used to translate the signed graph data to 2-dimension space where standard spectral clustering and k-means are both examined and are compared to existing state-of the art clustering methodologies on two sentiment-based datasets. We demonstrate that our approach successfully recovers all community labels on a highly modular dataset and performs on the level of the leading signed graph clustering techniques on a more complex network. Additionally, we demonstrate that status and influence, in combination with network data, can be used to detect and characterize anomalous outcomes in promotion networks.

CCS CONCEPTS

• **Networks** → **Social media networks**; • **Information systems** → **Clustering and classification**; • **Theory of computation** → **Social networks**.

KEYWORDS

Clustering, Community Detection, Social Networks, Signed Graphs

ACM Reference Format:
Maria Tomasso, Lucas J. Rusnak, and Jelena Tešić. 2022. Cluster Boosting and Data Discovery in Social Networks. In *Proceedings of ACM SAC Conference (SAC'22).* ACM, New York, NY, USA, Article 4, 3 pages. https://doi.org/10.1145/3477314.3507243

1 INTRODUCTION

In this paper, we go beyond statistical data analysis and look into sentiment-based considerations with a consensus-driven context for cluster and group formation. We examine three data sets - Highland Tribes [8], Sampson's Monastery [10], and Wikipedia Elections [5]. The Highland Tribes dataset describes agreeable and antagonistic relations between 16 tribes of the Eastern Central Highlands of New Guinea. Sampson's monastery data describes social relationships between eighteen novice monks in a New England monastery between 1966-1967 [10]. The wiki-Vote data set contains data on all Wikipedia adminship elections that took place before January 2008.

We introduce a new method for boosting ground-truth recovery based on [9] that introduced new metrics to quantify social impact by discovering minimal balancing states via a spanning-tree correction algorithm. We demonstrate that the status and influence metrics used to quantify the social status and influence of vertices perform on level with state-of-the art signed graph clustering methodologies, which out-perform standard spectral and k-means clustering. We considered three spectral clustering approaches from [4] (lap_none, lap_sym, lap_sym_sep), two balanced normalized cuts from [1] (BNC_none, BNC_sym), a novel spectral method solved via a generalized eigenproblem (SPONGE) [2] (SPONGE_sym), geometric means Laplacian (GM) [6], matrix power means Laplacian (SPM) [7], and we assess clustering success with the adjusted Rand index (ARI) by comparing generated community labels to known ground-truth values. When ground-truth is unknown we use our analysis to uncover Wikipedia Elections data anomalies. For future studies, we plan to overlay our approach with the other methods examined to determine if the state-of-the-art methods examined can also be boosted.

2 SOCIAL NETWORKS, SIGNED GRAPHS, AND BALANCE

Signed graphs are graphs where the edges are signed +1 or −1 if they represent an agreeable or an antagonistic relation, respectively. A signed graph Σ is *balanced* if the product of the signs of each cycle is positive. If Σ is not balanced then there exists sets of edges whose sign reversal produces a balanced signed graph each called a *balancing set.*

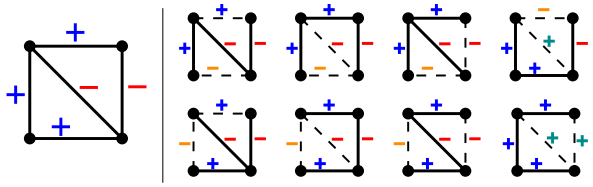

Figure 1: An unbalanced signed graph Σ (left) and its nearest balanced states (right). Spanning trees appear as solid edges and inherit their signs from the original signed graph, signs outside each tree that change from − to + are teal, while + to − are orange. Unchanged signs remain blue and red.

Rusnak et al. introduced new balance metrics to quantify social impact via the frustration cloud by discovering minimal balancing states via a spanning-tree correction algorithm in [9] where each

*The full version of the paper is available as `acmart.pdf` document

SAC'22, April 25 – April 29, 2022, Brno, Czech Republic

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507243

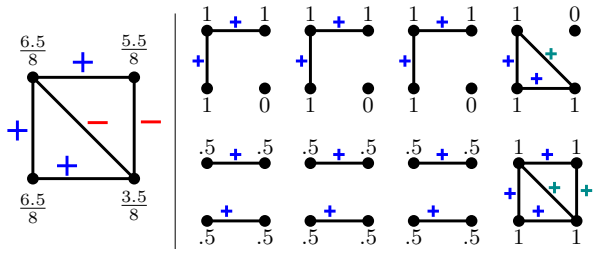

Figure 2: Vertex status measure (left) computed from consensus states (right). The deletion of negative edges results in the partition for consensus states from Fig. 1. Nodes are labeled by majority value for Σ (right).

fundamental cycle is balanced with respect to a given spanning tree to produce a balanced state. Figure 1 illustrates this process on a simple unbalanced signed graph Σ.

Figure 2 (right) illustrates the resulting consensus components obtained by the deletion of the negative edges from each balanced state as guaranteed by Harary's Theorem [3]. For each part of these partitions we assign each vertex a value of 1 if it belongs to the majority, and a value of 0 if it belongs to the minority; in the event of a tie all vertices are assigned a value of 0.5. These values are then averaged over all spanning trees used to provide a percent value called *status*.

The dual concept of status for edges is called *agreement*, and the average agreement of the edges around a given node is called the *influence* of the node. From [9] it is known that the that the status of a vertex is always greater than or equal to its influence, and the resulting status-influence cone exists in the first half of the first quadrant.

3 COMMUNITY DISCOVERY AND CLUSTER BOOSTING

The Highland Tribes data set has three ground-truth communities. Conventional community detection techniques are highly effective at recovering ground-truth labels on this dataset due to its inherent modularity: most positive edges occur within communities while most negative edges occur between communities. As shown in Figure 3 (top), most methods were often able to achieve 100% ground-truth recovery as measured by the adjusted Rand index (ARI).

Next we map the nodes from Highland Tribes into the status-influence cone in Figure 3 (bottom). Status and influence effectively separated the nodes based on ground-truth community membership. We exploit this separation by applying two classical clustering methods, spectral clustering (BCM_SC) and k-means clustering (KM_SC), to the status-influence coordinates before assessing the generated community labels using the ARI. This is shown in red in Figure 3 (top), where both BCM_SC and BCM_KM both succeeded in fully recovering the ground-truth labels.

Sampson's monastery describes social relationships between eighteen novice monks in a New England monastery between 1966-1967 [10]. Ground truth is present in this dataset because four social groups were identified throughout the study. For this experiment, we used data from the middle survey and assigned a single +1/-1 sentiment for each pair of monks.

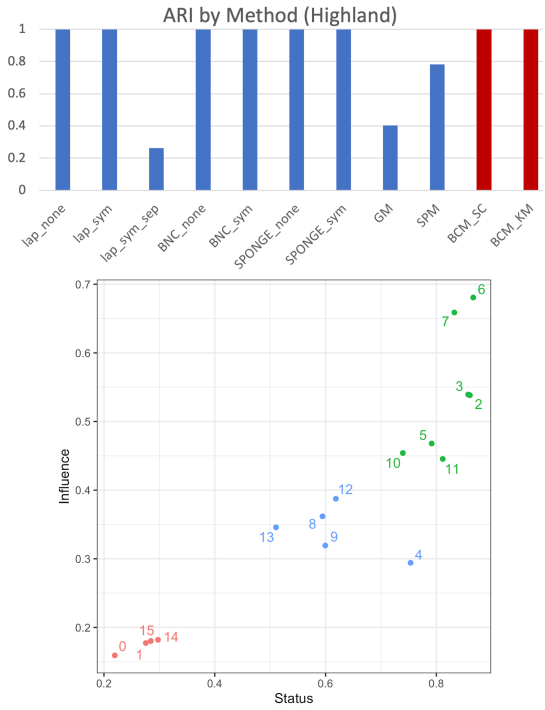

Figure 3: Top: The Highland Tribes Adjusted Rand Index (ARI) clusterability measure of signed spectral clustering methods on Highland tribes. Bottom: Ground truth communities for Highland Tribes in the status-influence space.

We repeated the clustering experiment conducted on Highland Tribes on Sampson and found that none of the signed clustering methods were able to capture the nuance of this underlying grouping, as illustrated in Figure 4 (top), with our method performing on the level with each of these models Figure 4 (bottom) depicts the difficulties associated to cluster detection in the status-influence cone.

4 WIKIPEDIA ELECTIONS ANALYSIS

The wiki-Vote data describes all Wikipedia adminship elections that took place before January 2008. In [9] status and influence were used to distinguish between adminship and promotion to adminship, and the status-influence cone for promoted users appears in Figure 5. Wikipedia administrators are editors who have been granted the ability to perform special tasks. Administrators are chosen through a *community review process* that seeks *consensus*, not a majority-rules decision.

Rather than predicting promotion, we are examining outliers through the lens of status and influence to characterize anomalous elections. All elections were examined with a logistic model on outcomes with RfA as a predictor. Mis-classified points were separated based on outcome and projected into status-influence space, as depicted in Figure 5.

The non-promoted logistic model outliers were traceable to three scenarios: mislabeled data in the original SNAP dataset when cross-checked via Wikipedia; the nomination was withdrawn by the

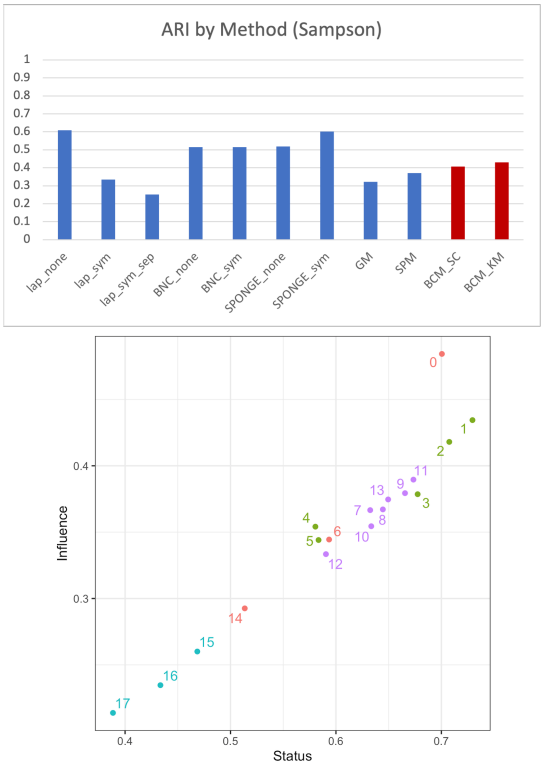

Figure 4: Top: ARI recovery score of ground-truth labels for Sampson. Bottom: Sampson ground truth view in status influence space.

candidate; consensus was not reached as determined by a bureaucrat. The promoted logistic model outliers were traceable to five scenarios: inexperience; judgement concerns; lack of need for admin tools; removed by arbitration; or a specific issue raised during the election cycle. These are shown in Figure 5.

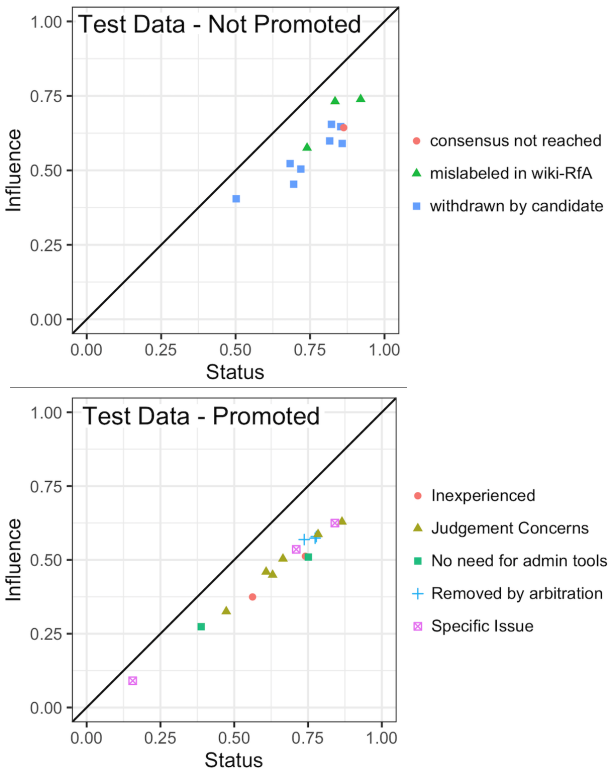

Figure 5: Top: wiki elections in status/influence space; winners are blue and losers are red. Mid: Non-promoted outliers. Bottom: Promoted outliers.

REFERENCES

[1] K-Y. Chiang, J.J. Whang, and I. 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*, page 615, Maui, Hawaii, USA, 2012. ACM Press.

[2] M. Cucuringu, P. Davies, A. Glielmo, and H. Tyagi. SPONGE: A generalized eigenproblem for clustering signed networks. *arXiv:1904.08575 [cs, math, stat]*, 2019.

[3] F. Harary. On the notion of balance of a signed graph. *Michigan Math. J.*, 2:143–146, 1953.

[4] 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. ACM, 2009.

[5] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[6] Pedro Mercado, Francesco Tudisco, and Matthias Hein. Clustering signed networks with the geometric mean of laplacians. *Advances in Neural Information Processing Systems*, 12 2016.

[7] Pedro Mercado, Francesco Tudisco, and Matthias Hein. Spectral Clustering of Signed Graphs via Matrix Power Means. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4526–4536, 2019.

[8] Kenneth Read. Cultures of the central highlands, new guinea. *Southwestern Journal of Anthropology*, 10(1):1–43, 1954.

[9] Lucas Rusnak and Jelena Tešić. Characterizing attitudinal network graphs through frustration cloud. *Data Mining and Knowledge Discovery*, 6, November 2021.

[10] S. Sampson. A novitiate in a period of change: An experimental and case study of relationships. *Ph.D. Thesis*, 1968.