Status correlation clustering in social networks

Dewan Ferdous Wahid
University of British Columbia

April 26, 2015

Introduction

Mutual attitude between members of a social group can be represented as a mixture of positive (friendly) and negative (hostile) interactions. Heider[1] first represented signed directed network to describe sentiment relation between people pertaining to a society. He also defined the balanced state of a social group based on the principle that "*friend of my friend is my friend* (Fig. 1 T_3)" whereas "*enemy of my enemy is my enemy* (Fig. 1 T_1)" and "*enemy of my enemy is my friend*". Catwright and Haray[2][3] formalized Heider's definition of a balanced state in graph-theoretic language, that a balanced signed directed network can be partitioned into mutually hostile subgroups, where the internal attitude between the members in these subgroups are supportive to each other. Davis[4] extended the definition for more than two clusterable subgroups. He proposed weak structural balance by eliminating the assumption "enemy of my enemy is my friend(Fig. 1 T_0)" as a balanced state.

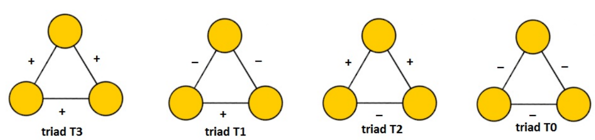

Figure 1: Heider's structural balance based on the number of positive edges. Triads with odd number of positive edges are *balanced* (T_3, T_1), and with even or no positive edges are *unbalanced* (T_2, T_0) [5].

A signed network is called *k-correlation-clusterable* if its vertices set can be partitioned into k sub-sets in such a way that the mutual relation between the members in each sub-set is positive and the mutual relation between two members from different subsets is negative. The *Correlation-clustering (CC)* problem, based on the balance theory, aims to find clusters in a social group with minimal clustering error. Negative edge inside a cluster/partition or a positive edge between two partitions results in a clustering error. Therefore, the optimal solution of the CC problem represents the degree of imbalance (i.e. number of clustering errors) in social networks [6]. From last decades, several works has been done (e.g. Doreian and Mrvar[7][8] [9], Figueiredo and Moura [6]) to analyze the correlation-clustering problem.

Apart from the balance theory, Leskovec et al.[5] proposed a new idea of *status* in social network analysis. They proposed that in a signed directed social network can be seen as a status-network of that society. In the status-network mutual attitude (friendly or hostile) between the members can be represented a the mutual status-relation. For example, a positive edge from vertex A to vertex B means *"B is friend of A"*, but it may also mean *"A thinks B has higher status than herself"*. Similarly, a negative edge form vertex A to vertex B means *"B is an enemy of A"*, but it may also mean *"A thinks B has lower status than herself"*. Authors used this idea to predict which type of links may evolve in a social network with propagation of the time.

In this work we consider a static picture of a directed signed network. We use the idea of *status theory* to define status balance and then formulate a novel type of correlation-clustering problem call *status-correlation-clustering (SCC)* problem based on status balance definition. We also formulate *status-degree-index* and give and idea on *status-clustering-coefficient* which we can as network measures in future analysis.

Why do we need status theory?

Balance theory was originally proposed for finding communities in undirected networks. Later on it has been widely using to analyze the structural balance of directed networks by simply ignoring the direction of edges[5][10]. The problem with the balance theory is that particular types of structures such as triangles with all positive edges are overrepresented, whereas triangles with one negative edge & two positive edges are widely underrepresented. These over and underrepresentation of certain type of triangles create biasness during predicting clusters in a social network. Leskovec et.al.[5] claims that the weak balance theory is generally consistent with the undirected networks whereas status theory shows more consistency with large scale signed directed networks (e.g. Epinions social network).

The status theory

Consider a situation in a social network when there is a directed positive edge between two members, from A to B. According to the status theory, it predicts that *"A regards B as having higher status"*. Also consider another positive edge from B to C, thus status theory predicts that *"B regards C as having higher status"* (Fig. 2). Now, what type of edge can we predict between A & C? The direction and sign of that edge depends on whether we look at it from A's point of view or C's point of view. Since A thinks B has a higher status than A, and B thinks C has higher status than B, then from A's point of view the edge from A to C is most likely a positive edge. That is, A, most likely, regards C of having a higher status than A. (Fig. 2(a)).

On the other hand if we consider from C's point of view, since B is regarded as having a higher status by A and again C is regarded as having a higher status by B. Then the most predictable edge from C to A is a negative edge. That is, C regards A, most likely, as having lower a status than C (Fig. 2(b)).

We can look another example from [5] for a better understanding of the status theory. Suppose all players in a soccer team are voting to select a team leader where any member

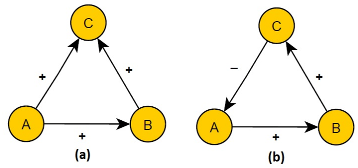

Figure 2: (a) From A's point of view, A regards B has a higher status and B regards C has a higher status. Then most likely, A regards C has a higher status. (b) C's point of view, B is evaluated as higher status vertex by A and C is evaluated as higher status vertex by B. Then most likely C regards A has a lower status than C.

can be a leader. A positive vote from A to B means, A regards B of having a better quality than herself. A negative vote from A to B means, A thinks she has a better quality than B. No vote means neutral to make any comments. Consider a situation where two players A and B both are positively evaluated by another player C. Now what type of response can we infer from A to B? For this we have to consider two different view points.

- From A's point of view, since A is positively evaluated by player C. Therefore, the evaluation that A gives B should be less likely to be positive than an evaluation received by B from a random team member.

- From B's point of view, since B is also positively evaluated by C. Therefore, the evaluation that A gives B should be more likely to be positive than an evaluation given by A to a random team member.

Therefore, we can say that the given status from a member A in a social group to another member B depends on the status received by A from other members.

Definition of status balance

Instead of regarding the dynamic evaluation of network, here we are considering a static snap-shot of a social network. We can formulate a definition of *balance status network* based on the Leskovec et.al.'s[5] *status theory.*

A network is balanced in status or simply status-balanced if it maintains a hierarchy of status among the members who form a triangle.

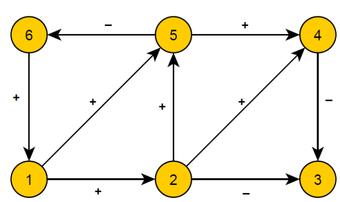

Figure 3: A status-balance signed directed network and a status hierarchy in maintained all triangles.

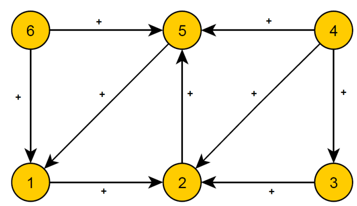

Figure 4: A status-imbalance signed directed network. There is a violation of status hierarchy in the triangle formed by vertices {1,2,5}.

In a network the status hierarchy is violated when a member having a lower status regards another member with higher status as negatively or a higher status member regards a lower status member as positively. More precisely, consider A thinks B has higher status than A, and B thinks C has more status than B. Then hierarchy of status is violated when A regards C of having lower status than A, or C regards A of having more status than C.

From the above notion we can categorize all triangles in a network into two groups: triangles where there are no violations of status and triangles where there are violations (e.g. Fig. 5). We can address the triangles with no violation and violation of status as *positive triangles* and *negative triangles* respectively.

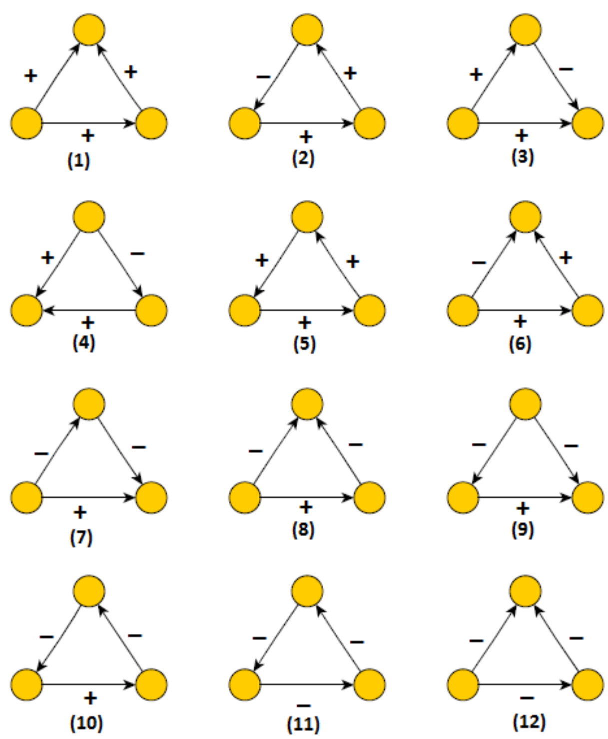

Figure 5: Triangles $1, 2, 3, 4, 8, 9, 10, 12$ are positive (i.e. status-balanced triangles) and triangles $5, 6, 7, 11$ are negative (i.e. status-imbalanced triangles).

k-status-clusterable network: Status-balance definition always looks for a status hierarchy inside a triangle. But in a network there may exist an edge which does not participate to any positive or negative triangle. Therefore, to formulate the definition of a network which is *k-status-clusterable* we use two notions : *k-correlation-clusterable network* and *status-balance network*.

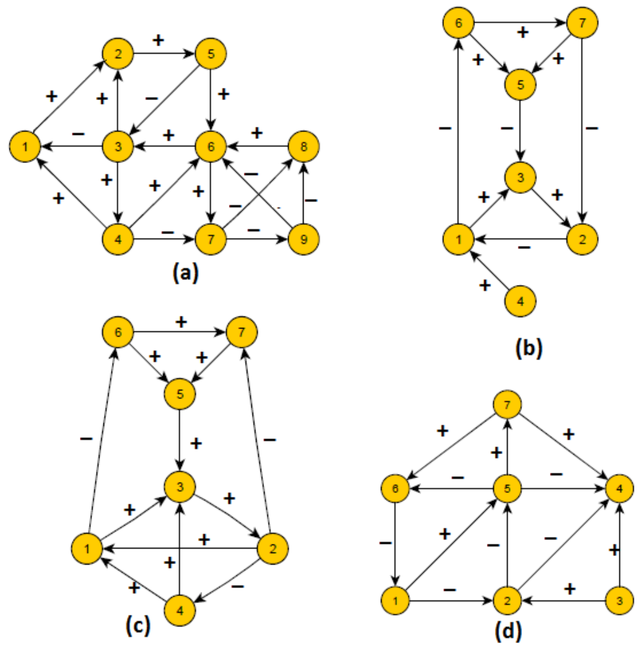

Figure 6: (a) *k-status-clusterable* with clusters $\{1,2,3,5\},\{4\},\{6,7,8,9\}$ and $k=3$. (b) *k-status-clusterable* with clusters $\{1,2,3,4\},\{5,6,7\}$ and $k=2$. (c) not *k-status-clusterable*. Gives status-cluster error 4 with partitions $\{1,2,3,4\},\{5,6,7\}$. Since vertices $\{1,2,3\}$ forms a negative triangle and are inside the same partition. Also edge $(5,3)\in E^+$ but 5 & 3 are in different partitions. (d) not *k-status-clusterable*. Gives status-cluster error 3 with the partitions $\{1,5,6,7\},\{2,3,4\}$. The cluster error comes from the positive triangle formed by the vertices $\{2,4,5\}$ which exist between two partitions.

Let $G=(V,E,s)$ be a signed directed network, where $V=\{1,2,...,n\}$ is the set of all vertices and E is the set of all edges, and $s:E\rightarrow\{+,-\}$ is a function that assigns a sign to each edge $e\in E$. That is an edge $e\in E$ is *positive* if $s(e)=+$, and *negative* if $s(e)=-$. Let E^+ and E^- be the sets of all positive and negative edges respectively. Let $t_{ij}^p: V\rightarrow\{-1,0,1\}$ be a fuction representing the vertices $i,j,p\in V$ forming either a positive or negative triangle or no triangle; i.e.

$$t_{ij}^p=\begin{cases} -1 & : (i,j)\in E \text{ forms a negative triangle with any } p\in V,\\ 0 & : (i,j)\in E \text{ does not form any triangle with any } p\in V,\\ 1 & : (i,j)\in E \text{ forms a positive triangle with any } p\in V.\end{cases}$$

Then we can formulate the definition as follows:

A signed directed network $G = (V, E, s)$ is said to be *k-status-clusterable* if its vertex set V can be partitioned into subsets $V_1, V_2, ..., V_l$ where $l \leq k$ such that, vertices $i, j, p \in V$ with $t_{ij}^p = 1$ are in the same cluster and vertices $i, j, p \in V$ with $t_{ij}^p = -1$ are in two or more different clusters. An edge $(i, j) \in E$ with $t_{ij}^p = 0; \forall p \in V$, if $(i, j) \in E^+$ then i & j are in the same cluster and if $(i, j) \in E^+$ then i & j are in different clusters.

The limitation of the above definition is that, it does not give any idea about the status-violation when there is no triangle in network or if the status-violation occurs in a long cycle. For example in Fig. 7. In future we have generalize this status-balance definition.

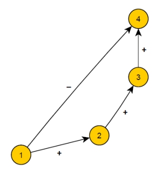

Figure 7: Status balance definition does not give any direction for the graph when there is no triangle.

Status correlation clustering problem

According to the definition of *k-status-clusterable* a status-balanced signed directed network doest not allow any negative triangle inside a cluster and any positive triangle between two clusters. Moreover if an edge $(i, j) \in E^+$ does not participate to form a triangle and if vertices i & j are in different clusters then there is a cluster error. Also if an edge $(i, j) \in E^-$ does not participate in forming a triangle and if i & j are in same cluster then it is also an error. Consider a partition $P = \{V_1, ..., V_2\}; l \leq k$ of vertices set V. Let w_{ij} be a nonnegative edge weight associated with each edge $(i, j) \in E$ and $i, j \in V$. Then the status-clustering error due to the partition P can be written as:

$$SCC(P) = \sum_{\substack{(i,j)\in E;\ i\neq j\neq p;\\ i,j,p\in V;\ t_{ij}^{p}=-1;\\ i,j,p \text{ are in same cluster}}} w_{ij} + \sum_{\substack{(i,j)\in E;\ i\neq j\neq p;\\ i,j,p\in V;\ t_{ij}^{p}=1;\\ i,j,p \text{ are in different clusters}}} w_{ij} + \sum_{\substack{(i,j)\in E^-;\ i\neq j\neq p;\\ i,j,p\in V;\ t_{ij}^{p}=0;\\ i,j,p \text{ are in same cluster}}} w_{ij} + \sum_{\substack{(i,j)\in E^+;\ i\neq j\neq p;\\ i,j,p\in V;\ t_{ij}^{p}=0;\\ i,j,p \text{ are in different cluster}}} w_{ij} \quad ; \text{for all } p \in V$$

Therefore we can define the status-correlation-clustering problem as follows:

Problem: For a given signed weighted directed network $G = (V, E, s)$ the status-correlation-clustering problem is to find a partition P such that the status-clustering error due to this partition, i.e. $SCC(P)$, is minimized.

Integer linear programing formulation of SCC problem

The classical integer linear programming (ILP) formulation for correlation-clustering problem was proposed by Demaine et al.(2006)[11]. In this section, we use this idea to formulate an ILP model for status-correlation-clustering problem described in the previous section.

We assign a binary decision variable y_{ij}^p for each vertices $i, j, p \in V$, $i \neq j \neq p$ as follows.

$$x_{ij}^p = \begin{cases} 0 : \text{if vertex } i,\ j \text{ and } p \text{ form a triangle and are in the same cluster,} \\ 1 : \text{otherwise.} \end{cases}$$

We also another assign a binary decision variable y_{ij} to each pair of vertices $i, j \in V$, $i \neq j$ and define as follows.

$$y_{ij} = \begin{cases} 0 : \text{if vertex } i \text{ and } j \text{ are in the same cluser,} \\ 1 : \text{otherwise.} \end{cases}$$

Then the objective function of SCC problem can be written as,

$$\min \sum_{(i,j)\in E;\ i\neq j\neq p;\ t_{ij}^p=-1} w_{ij}(1-x_{ij}^p) + \sum_{(i,j)\in E;\ i\neq j\neq p;\ t_{ij}^p=1} w_{ij}x_{ij}^p + \sum_{(i,j)\in E^-;\ i\neq j;\ t_{ij}^p=0} \alpha w_{ij}(1-y_{ij}) + \sum_{(i,j)\in E^+;\ i\neq j;\ t_{ij}^p=0} \beta w_{ij}y_{ij}. \quad (1)$$

Here we add two parameters α and β to the last two terms of the objective function. We can use these two parameters to control the effects of an edge on the status-clustering error, which does not participate to form a triangle. For the rest of the part of this report we consider $\alpha = \beta = 1$.

The first constraint is that, when $x_{ij}^p = 0$ then the vertices i, j and p are in the same cluster. This constraint can be formulated as,

$$0 \leq y_{ij} + y_{ip} + y_{pj} - x_{ij}^p \leq 2, \quad (2)$$
$$x_{ij}^p - y_{ij} \geq 0, \quad (3)$$
$$x_{ij}^p - y_{ip} \geq 0, \quad (4)$$
$$x_{ij}^p - y_{pj} \geq 0, \quad (5)$$
$$x_{ij}^p = \{0,1\}, \quad y_{ij} = \{0,1\}. \quad (6)$$

Also all vertices either are in the same cluster or in different clusters. This can be defined as,

$$y_{ip} + y_{pj} \geq y_{ij}, \quad (7)$$
$$y_{ij} = y_{ji}. \quad (8)$$

The objective function [Equation 1] with the constraints (2)-(8) gives the ILP formulation of SCC problems.

Implementation of the ILP model

We have implemented this ILP model in Java using JUNG graph package and IBM ILOG Cplex version 12.6. We have tested this model on few small signed directed networks which are given in *Appendix*.

Status network measures

Clustering coefficient

We use network measures to analyze the properties of a network by numerical representations. Clustering coefficient for undirected networks, which was proposed by Watts et al. [12], is one of the efficient measures that represents the likelihood of connectivity in the neighborhood of a vertex. In social network interpretation, clustering coefficient represents the probability that two member in a social group with a common friend are also friends. It can also be considered as the index of redundancy of a vertex in a network. Later Onnela et al.[13], Zhang et al. [14] and Saramaki et al.[15] generalized the clustering coefficient for weighted networks.

Recently, Costantini et al.[16] proposed a generalization of clustering coefficients for signed weighted undirected networks. Suppose the positive and negative edges represent the mutual friendship and antagonist relation between two members in a social group. We know clustering coefficient indicates how redundant is a vertex i to pass an information between two of its neighbors j & k. If the passing information between vertices j & k through direct path $j-k$ and indirect path $j-i-k$ are same then i is more likely redundant between j & k, i.e. i has high clustering coefficient. On the other hand, if the information through direct and indirect path are different then i has an important roll to pass information between vertices j & k. In this case, i has low clustering coefficient.

Therefore we can summarize that, the signed clustering coefficient of a vertex i is high if two of it neighbor with same sign relation has a positive relation or the pair of neighbors that have opposite sign relation with i have a negative relation (e.g. Fig. 8).

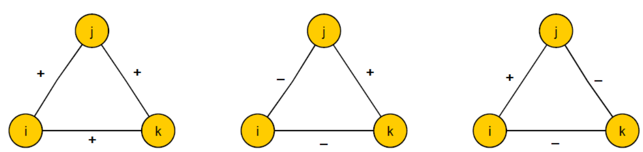

Figure 8: Vertex i has high clustering coefficient.

Moreover, i has low signed clustering coefficient if the mutual relation between two of its neighbors that have same sign connections with i is negative or if the pair of its neighbors that have opposite sign relation with i has a positive relation (e.g. Fig. 9).

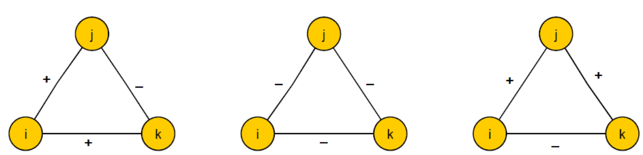

Figure 9: Vertex i has low clustering coefficient.

Status clustering coefficient: In a status-network the sign of outer edge (i.e. the attitudes towards other members) from a vertex depends on the sign of inner edges (i.e. the attitudes received from other member). Also the connection between two neighbors of a vertex i depends on their points view. This connection most often changes with the changes in view points.

We know that clustering-coefficient represents the index of redundancy of a vertex to pass information between its neighbors. By using this notion, we can consider the *status-coefficient* of a vertex $i \in V$ as, how important is that vertex to get information about the status of one of its neighbor j from another neighbor k. That is, if the status information is same in the direct $k-j$ and indirect path $k-i-j$ then i has high status-clustering coefficient. And if the information is different then i has low clustering coefficient. We therefore notice that any vertex that participates in the formation of a positive triangle (e.g. Fig. 5) has a high clustering coefficient and the one that participates in the formation of a negative triangle has a low clustering coefficient (e.g. triangles 5-7 & 11 in Fig. 5).

In future we shall work upon the mathematical formulation of the *status-clustering-coefficient* of a vertex in a signed directed network.

Status degree-index

According to the status theory, a positive edge from vertex i to vertex j means *"i regards j of having a higher status than itself"* and a negative edge from vertex i to j means *"i regards j of having a lower status than itself"*. Let w_{ij} be the weight of the edge from i to j. Now we consider the following cases,

- if $(i, j) \in E^+$ the status of the vertices i and j are decreased and increased by w_{ij} respectively.
- if $(i, j) \in E^-$ the status of the vertices i and j are increased and decreased by w_{ij} respectively.

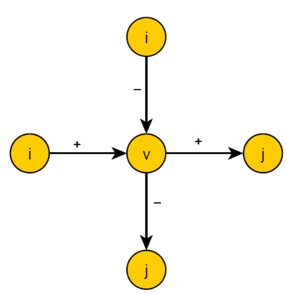

Figure 10: Status degree-index of vertex v.

Let s_{ij} be the sign of the edge $(i, j) \in E$. Therefore the *status degree-index* of any vertex v can be defined as,

$$SDI(v) = \frac{\sum_{i \in V} s_{iv} w_{iv}}{\sum_{j \in V} |s_{vj} w_{vj}|}$$

The value of $SDI(v)$ lies between $[-1, 1]$. In a social network the status degree-index of a vertex can be interpreted as the degree of confidence as a social leader, i.e. $SDI(v) = 1$ and $SDI(v) = -1$ means vertex v is the most and the least confident person to be a leader respectively.

Conclusions and future works

We have formulated the definition of a balance in a directed signed network induced by a social group, based on status attitude between two members of that society. Then we describe a status-correlation-clustering (SCC) problem and model an integer linear programing formulation.

These definitions and model are still trivial in sense, as they do not give any direction for the following cases.

- We only consider the status violation within a triangle. According to the definition of the status-balance, a network is balanced if there is a hierarchy of status among the members that participate a triangle. But if the network is completely triangle free (e.g. trees) or if there is a status violation in a long cycle (e.g. Fig. 7), then our model does not work.
- We are yet to prove any feasibility and time complexity theorem behind the model. In future we shall prove the following theorems.
  - **Theorem 1 (existence of feasible solutions):** For a status-balanced signed directed network there always exist at least one partition of vertices where each of the subsets of vertices are status-balanced.
  - **Theorem 2 (time complexity):** SCC problem is an NP-hard (it is a conjecture).
- We did not consider the case if there exist edges in both directions between two vertices.

We have not proposed any mathematical formulation to measure the status-clustering-coefficient of a vertex in signed directed network. Here we have proposed another simple network measure *status degree-index*, but we are yet to explore the application of this measure. In future we shall work on resolving these issues and shall test the model for large scale networks. Also we shall work on developing network measures to analyze the properties of status-networks.

Appendix

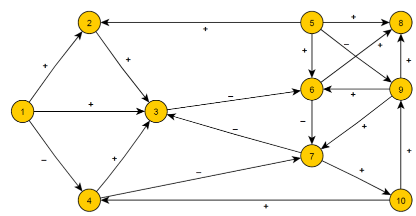

Figure 11: Clustering error is 4.0 and the clusters: $\{1,2,3,4,10\}, \{5,6,7,8,9\}$. The error comes from the positive edge $(2,5)$ between two clusters and the negative triangle $(7,9,10)$ inside a cluster.

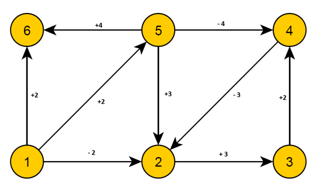

Figure 12: Clustering error is 0.0 and the clusters:$\{1,5,6\}, \{2,3,4\}$. Therefore this network is status-balanced.

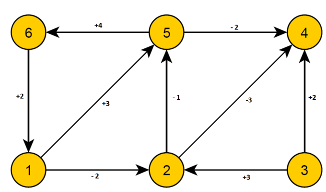

Figure 13: Clustering error is 0.0 and the clusters:$\{1\}, \{2,3,4,5,6\}$. Therefore this network is status-balanced.

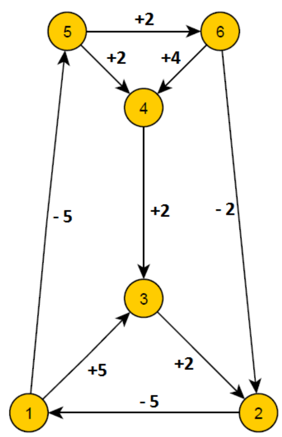

Figure 14: Clustering error is 2.0 and the clusters: $\{1\}, \{2, 3, 4, 5, 6\}$. This error comes from the positive edge $(4, 3)$ between two clusters.

References

[1] F. Heider, "Attitudes and cognitive organization," *The Journal of psychology*, vol. 21, no. 1, pp. 107–112, 1946. [Online]. Available: http://psychclassics.yorku.ca/Heider/attitudes.htm

[2] D. Cartwright and F. Harary, "Structural balance: a generalization of heider's theory." *Psychological review*, vol. 63, no. 5, p. 277, 1956. [Online]. Available: http://dx.doi.org/10.1037/h0046049

[3] F. Harary, "On the measurement of structural balance," *Behavioral Science*, vol. 4, no. 4, pp. 316–323, 1959. [Online]. Available: http://dx.doi.org/10.1002/bs.3830040405

[4] J. A. Davis, "Clustering and structural balance in graphs," *Social networks. A developing paradigm*, pp. 27–34, 1977. [Online]. Available: http://dx.doi.org/10.1177/001872676702000206

[5] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2010, pp. 1361–1370. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753532

[6] R. Figueiredo and G. Moura, "Mixed integer programming formulations for clustering problems related to structural balance," *Social Networks*, vol. 35, no. 4, pp. 639–651, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.socnet.2013.09.002

[7] P. Doreian and A. Mrvar, "A partitioning approach to structural balance," *Social networks*, vol. 18, no. 2, pp. 149–168, 1996. [Online]. Available: http://dx.doi.org/10.1016/0378-8733(95)00259-6

[8] P. Doreian, "A multiple indicator approach to blockmodeling signed networks," *Social Networks*, vol. 30, no. 3, pp. 247–258, 2008. [Online]. Available: http://doi:10.1016/j.socnet.2008.03.005

[9] P. Doreian and A. Mrvar, "Partitioning signed social networks," *Social Networks*, vol. 31, no. 1, pp. 1–11, 2009. [Online]. Available: http://doi:10.1016/j.socnet.2008.08.001

[10] S. Wasserman, *Social network analysis: Methods and applications.* Cambridge university press, 1994, vol. 8.

[11] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica, "Correlation clustering in general weighted graphs," *Theoretical Computer Science*, vol. 361, no. 2, pp. 172–187, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.tcs.2006.05.008

[12] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998. [Online]. Available: http://dx.doi.org/10.1038/30918

[13] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, "Intensity and coherence of motifs in weighted complex networks," *Phys. Rev. E*, vol. 71, p. 065103, Jun 2005. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.71.065103

[14] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, pp. 1–7, Aug 2005. [Online]. Available: http://dx.doi.org/10.2202/1544-6115.1128

[15] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész, "Generalizations of the clustering coefficient to weighted complex networks," *Phys. Rev. E*, vol. 75, p. 027105, Feb 2007. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.75.027105

[16] G. Costantini and M. Perugini, "Generalization of clustering coefficients to signed correlation networks," *PloS one*, vol. 9, no. 2, p. e88669, 2014. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0088669