Threshold Models on Signed Networks

Candidate Number: 649440

Submitted to the Mathematical Institute, University of Oxford

The principal aim of this project is to construct preliminary threshold models on signed networks and test their robustness on toy datasets. In particular, we will use the example of information spreading through social networks as our key motivation, incorporating into our model the importance of the global structure of such networks. The initial work of Ehsani et al.(1) into this topic will be discussed and refined to motivate a new model. We will explore the notion of structural balance within signed networks, the subject of large amounts of existing literature, and results from balance theory will allow us to partition the network into individual communities. The key idea from which our model will spawn is the tight-knit nature of these communities in social networks – information will spread within individual communities much faster than between two communities with largely antagonistic relations. Ultimately, this will lead us to develop metrics that quantify the strength of these relations and parameterise our model.

Signed Networks | Threshold Models | Structural Balance

The use of signed networks, in which a positive or negative orientation of each edge must be specified, is an incredibly powerful tool to describe real-world data, providing us with a great deal more information than standard networks. Indeed, they often have a very natural physical motivation, especially in the context of friendly and antagonistic relations in social groups. In spite of this, signed data is often ignored in the analysis even when it is available and may further our understanding, for example in (2). The example dataset here from Everett and Borgatti's paper is a classical signed social network of Gahuku-Gama intertribal relations in New Guinea, originating from a 1954 anthropological study (3) - a dataset we will study extensively later to test our model.

In the aim of notational simplicity, we shall refer to networks without specified edge-signs as *simple networks*. Further, positive connections in signed graphs will be represented by a solid line and negative connections by a dashed line (in some cases positive and negative edges will be coloured green and red, respectively, to further highlight their orientation).

The lack of attention paid to spreading processes on signed networks thus far can perhaps be explained by the difficulty and lack of intuition associated with the problem. The flow of information down edges can often be physically motivated in simple networks, while in signed networks the concept of negative edges blocking flow is more difficult to deal with. The natural intuition of simply preventing a flow of information along these negative edges does not yield us any progress, since this would be to treat a negative edge in identical fashion to no edge. Again, this makes sense physically - if we look only at a single node and their connections then there is little difference in the rate of passing information between a neutral connection (no edge) and a negative connection. This already hints that we need to consider the global structure of the network rather than simply an egocentric visualisation of each node - there are more complicated structures, communities, embedded within the network which impact the flow.

For example, consider figure 1, node A is less likely to pass secretive information to node B, given B's propensity to pass it to node C, who has an antagonistic connection with our original node A.

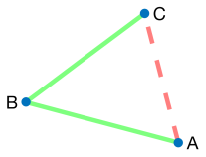

Fig. 1. Example of an unbalanced network

This is precisely the concept of structural balance and it is here we will start out investigation to gain a better understanding of the global structure of our signed networks. In the following section we will cover only the key results and intuition which will prove useful for us, omitting most proofs - extensive references are provided for the reader to explore the subject in more depth at their leisure.

Structural Balance

Definition 1. *A closed loop (cycle) within a network is* ***(structurally) balanced*** *if it does not have precisely 1 negative edge.*

Definition 2. *A network is* ***(structurally) balanced*** *if it contains no unbalanced cycles.*

Note that these definitions differ slightly from those often quoted, for example in Newman (4). We have chosen this particular definition to allow for an arbitrary number of communities in our node partition*.

* Other sources define a cycle to be balanced if the product of edge signs around the cycle is positive.

Significance Statement

The topic of spreading processes on networks has been studied in great detail with broad physical applications. In particular, a large volume of literature is devoted to studying infection models and information passing through social networks. Signed networks, with edges allowed to be either positive or negative, are largely neglected in such study, despite presenting a natural representation of social situations. This project ventures into somewhat unexplored territory in an attempt to construct a new method to model these processes, taking extensive motivation from the spreading of information across social groups. Certainly, more questions will be asked than answered, but hopefully the following should provide a launching pad for further research.

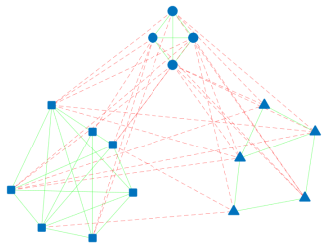

Fig. 2. Gahuku-Gama intertribe relations dataset(3): 3-partition demonstrates structural balance on partial data

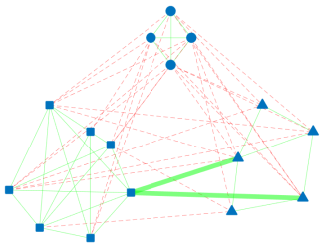

Fig. 3. Gahuku-Gama intertribe relations dataset(3): highlighted edges disrupt structural balance

Theorem 3. (Davis (5), 1967) *The nodes of a balanced signed network can be partitioned into two or more subsets (clusters) such that:*
i) Every positive edge joins nodes in the same cluster.
ii) Every negative edge joins nodes of different clusters.

Definition 4. *A network is* ***k-balanced*** *if it can be partitioned into k clusters in the setting of* Theorem 3.

Definition 5. *A* ***cluster*** *is a set of the k-partition such that among the nodes in the cluster there are only positive edges and any edges to the remainder of the network are negative.*

In practice, we will look for partitions that minimise the number of edges that disrupt the structural balance, and in this case we informally refer to the subsets of nodes as **communities**.

For example, figure 2 shows the partition of nodes into three separate clusters (represented by different node marker shapes) from an incomplete version of the Gahuku-Gama dataset (3) with the ties disrupting this structure removed.

A key assumption of balance theory is that real-world networks will tends towards structural balance, all other configurations are unstable. One can think of a community in the partition as a friendship group amongst whom there are only positive connections. This group then has only anatagonistic connections to all rival groups. Any edge that doesn't adhere to this structure can cause instability (for instance, if there is a disliked member within the friendship group, or if a member starts to develops stronger friendships with another group).

Experimentally, however, we often find examples of networks without structural balance, as in the full Gahuku-Gama tribe dataset, figure 3. We will find, in fact, that networks without full structural balance will exhibit much more interesting dynamics for our processes. As noted in Ehsani et al.(1), communities that consist only of internal positive connections and external negative connections serve to block any spreading process through the network. This is clearly apparent in our social network, the lack of positive edges between two communities prevent information from ever passing from one to the other. As such, the process is contained within a cluster containing no negative connections, thus reducing the dynamics to those of the well-studied simple networks.

Threshold Models

Before building a threshold-style model for our signed network, we shall first motivate it by studying basic threshold models on simple networks and Ehsani's (1) initial model on signed networks.

On a simple network, and in the most basic form, a threshold model acts on nodes which can exist in one of two states, let's say A and B. In the context of epidemiology, suppose that state A represents the susceptibles, those who are yet to be infected, and state B represents the infected population. At an initial time, $t = t_0$, an initial seed network of nodes are placed in state B, all remaining nodes are left as susceptibles. At subsequent timesteps we consider each individual susceptible node, changing it's state to B only if the proportion of infected neighbours is greater than a certain threshold, q.

That is, at time t, node x (in state A, susceptible) is infected if and only if:

$$q < \frac{\sum_{y \in \Gamma(x)} \mathbb{1}_{\{\text{y in state B, infected, at time t}\}}}{k(x)}$$

with $\Gamma(x)$ the neighbourhood of x, $k(x) = |\Gamma(x)|$ the degree of x.

Definition 6. *A* ***complete cascade*** *occurs if the process results in every node becoming infected.*

It is trivial to note that the number of nodes in state B increases monotonically at each time step until either the process is halted, with some nodes remaining in state A, or a complete cascade occurs.

Definition 7. *A* ***community of density p*** *is a set of nodes such that each node in the set has at least a fraction p of it's neighbours also in the set.*

Theorem 8. Conditions for Complete Cascade. (Easley et al.(6), 2010) *An initial seed network results in complete cascade if and only if the remainder of the network contains no community of density greater than* $1 - q$.

From this theorem we see that dense communities act as blocks to the cascade process, an analogous result to our earlier assertion that structural balance inhibits the spreading process in signed networks.

Ehsani's Threshold Model. Ehsani et al.(1) naturally extend this model to incorporate signed networks. The principal idea is that a node, x, will update it's state based only on the neighbours; neighbours with positive connections will pull x towards their state, while negatively-connected neighbours will push x away from their state.

We still assume two possible states for each node, with nodes only switching from state A to state B.

For any node x, in state A, define:

n_1 = number of positive neighbours in state A
n_2 = number of positive neighbours in state B
n_3 = number of negative neighbours in state A
n_4 = number of negative neighbours in state B

There are, therefore, $n_2 + n_3$ neighbours pulling x towards state B, and $n_1 + n_4$ neighbours pulling x towards state A.

Hence, at time t, node x switches to state B if and only if:

$$q < \frac{n_2 + n_3}{n_1 + n_2 + n_3 + n_4}$$

for threshold q.

Construction of a New Community-Based Model

Ehsani's model is both simple and well represents local dynamics with each node's state having a strong dependence on both the positive and negative connections. However, the weakness of such an approach is that it is egocentric - it focuses only on local aspects of the network and ignores the global structure. Empirical evidence would seem to suggest that signed networks have a much greater dependence on this global structure, as explained with figure 1. It may be that such behaviour is contained in the time-series approach of Ehsani's method, with global structures having carry-through effects to node activation, but such an assertion does not seem a trivial statement.

In order to decompose networks into tightly-connected communities which we can analyse independently, we use a similar approach to Doreian's, from his book (7) and 2009 paper (8). For this, we count all edges in a partition that are not consistent with structural balance.

Definition 9. *The penalty function,* $\boldsymbol{Q(C)}$*, for a partition of nodes,* C*, is defined such that:*

$$Q(C) = \alpha N + (1 - \alpha)P$$

N *= the total number of negative ties that are contained within a single community*
P *= is the total number of positive ties that span two different communities*
$\alpha \in (0, 1)$ is a weighting parameter

Note that increasing the value of our parameter, α, places greater weight on inconsistent negative edges within communities.

For our chosen parameter value, the optimal community structure will be given by the partition $C^* = \text{argmin}_C(Q(C))$.

For our method, however, all that we require is a partition of the network into communities, and this need not be achieved by solving explicitly for C^*. Instead, there exists a wealth of research into community detection with signed networks and any method may be used to obtain our initial partition, see (9), (10), (11). In particular, it may be that the community structure of the network is suggested by the empirical observations or otherwise pre-specified.

For this reason we shall not focus on the different approaches available, the best method will be determined by the particular scenario. Doreian's paper (8) does, though, detail an algorithm for which we are able to optimise our penalty function Q. In all of our examples below we shall pre-specify the community structure.

Definition 10.
i) N_i = *total number of negative ties within community* i
ii) $P_{i,j}$ = *total number of positive ties between communities* i,j
iii) $Q_i = \alpha N_i + \frac{1}{2}(1 - \alpha) \sum_j P_{i,j}$

Note now that $Q = \sum_i Q_i$, so we refer to Q_i as the ***strength of community*** $\boldsymbol{i}$.

Return now to our initial motivation, the spread of rumours through a social network. We have assumed that these rumours are highly sensitive to the global community structure of the network so that it is necessary to consider each community separately. We have dealt with, or otherwise avoided, the issue of determining communities and have now defined a metric to quantify the strength of each individual community.

Now consider one individual in the network in possession of the rumour. We assume that they have a strong understanding of the global structure of the network and this influences whether the rumour is passed at each time step. We assume every node in a community to be equivalent, so all that is left to define is the passing criterion for each community.

It is here that we also take a slight detour from traditional threshold model approaches, invoking a probabilistic element: **a node in possession of the rumour,** $\boldsymbol{x}$**, considers each positive connection,** $\boldsymbol{y}$**, independently and passes on the rumour with probability** $\boldsymbol{z_{x,y}}$.

Definition 11. $z_{x,y}$ *is the transfer probability for nodes* x,y. $\boldsymbol{Z} = (z_{x,y})_{x,y}$ *is the transfer probability matrix.*

Note that $\mathbf{Z}$ has the following properties:

- $z_{x,y}$ depends on x and y only through their communities
- $\mathbf{Z}$ is symmetric: $z_{x,y} = z_{y,x}\ \forall\ x, y$
- $z_{x,x}$ is unspecified and may be defined as is convenient

For notational simplicity, and because of the first property, we will write $z_{i,j}$ for the transfer probability between any

node x in community i and node y in community j.

In specifying our transfer probabilities we must now distinguish between cases $i = j$ and $i \neq j$.

Case I: $i = j$.
This is our transfer probability within community i. We desire the following properties:

- $z_{i,i}$ inversely proportional to Q_i: the transfer probability decreases for weaker communities
- $z_{i,i}$ directly proportional to the number of positive ties within the community: we have already incorporated this property by implementing independent trials with each positive connection for the passing of the rumour

Definition 12. $z_{i,i} = \mu \frac{\min_j(Q_j)}{Q_i}$ [†]

Case II: $i \neq j$.
This is our transfer probability between distinct communities, i and j. We desire the following properties:

- $z_{i,j}$ should take low values for any i,j - there is low probability of passing on information to 'rival' communities
- $z_{i,j}$ inversely proportional to the number of negative ties between the two communities: a greater rivalry between two communities should decrease the the transfer probability
- $z_{i,j}$ directly proportional to the number of positive ties between nodes: this property is already incorporated, as above, with the independent trials for each node and each of it's connections

Definition 13. $z_{i,j} = \epsilon \frac{1}{1+N_{i,j}}$ [‡]

With these definitions in place, we are now able to rigourously define the model through an algorithm.

Parameter Values. The definition of our transfer probability matrix, $\mathbf{Z}$, is parameterised by two constants, μ and ϵ, which we may specify or determine empirically. This allows us two degrees of freedom within our model, which may in fact be reduced to one by suitable rescaling of time. This one remaining degree of freedom is expected and parameterises the balance of the processes within and between communities.

More generally:

- μ: average probability of passing on information in the strongest community
- ϵ: average probability of passing on information between nodes in two different communities in the limit as the negative connections between them tends to 0

Under normal conditions we expect to see $\mu >> \epsilon$, corresponding to a greater rate of transfer of information within dense clusters than across rival communities.

[†] In the case that $\min_j(Q_j) = 0$, set $z_{i,i} = \mu$ for $i = \text{argmin}_j(Q_j)$ and take $\min_j(Q_j)$ to be half the smallest permissible value under α.

[‡] The 1 in the denominator ensures a well-defined probability in $[0, 1]$.

Algorithm: Community-Based Spreading Process

Our model will take any given network, with a specified community partition and initialisation, and simulate the spreading process described above on a two-state system. Nodes in state B will be referred to as 'active' and transferring state from A to B will be referred to as 'activation'.

Input:

- Network specified by adjacency matrix, $\mathbf{A}$. Note that this network need not be symmetric, our model incorporates directed networks.
- Node partition, C, defining communities in the network
- Initial seed network of activated nodes, S
- A run-time for the algorithm, the number of iterations t_{final} [§]

Algorithm:

1. Calculate transfer probability matrix, $\mathbf{Z}$, from the data
2. Initialise the system: assign all nodes in S to state B, all remaining nodes to state A. Set $t = 0$
3. $t = t + 1$
4. For each active node, x, consider each positive connection, y, in turn and independently activate y with probability $z_{x,y}$
5. If $t = t_{\text{final}}$ then exit. Else, return to step 3.

Crucially, this model incorporates all of the properties we desire. Namely, it is cheap and efficient to run and allows information to flow down positive connections whilst factoring in the effect the global community structure has on this flow.

Note that the relationship of having both positive and negative ties has been incorporated with the probabilistic measure. In contrast to Ehsani's (1) and the standard threshold model, there is no fixed criterion for activation, and this allows us to include more global properties of the network in Z. If we were to include conditions on global properties of the network into a fixed, non-probabilistic model then the spreading process would take much larger leaps because of the inter-dependency of different nodes' criterion.

Further, this probabilistic model can also be physically motivated from our rumour example. A node is likely to pass this information not based on a strict criterion but instead with inherent randomness. The probability at each turn of passing the rumour would then be dependent on structural properties of the network.

Simulations and Results

Let us first test this new model on a small dataset which we understand the structure of well. This will help to test the robustness of the model - we can compare the results obtained to what one would expect with prior structural knowledge.

[§] In all our simulations we shall run the process to completion: $t_{\text{final}} = \infty$

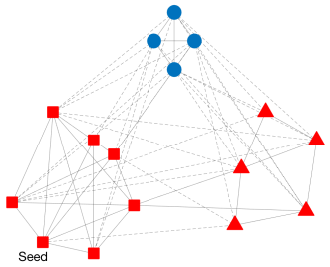

Fig. 4. Activation states in simulation of Gahuku-Game data. $\mu = 0.5$, $\epsilon = 0.1$. Nodes in state B (activated) are shown in red, blue nodes are in state A.

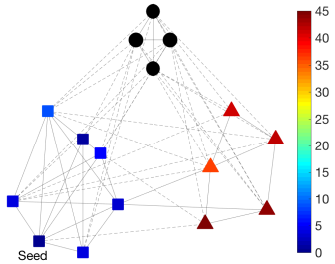

Fig. 5. Activation times in simulation of Gahuku-Game data. $\mu = 0.5$, $\epsilon = 0.1$.¶

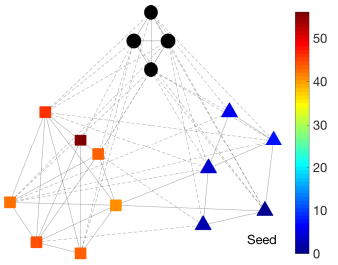

Fig. 6. Alternative seed node: $\mu = 0.5$, $\epsilon = 0.1$.¶

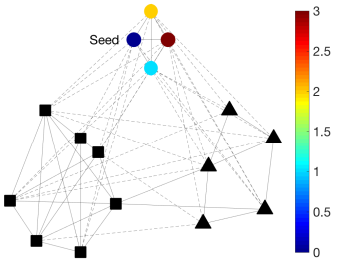

Fig. 7. Alternative seed node (within closed cluster): $\mu = 0.5$, $\epsilon = 0.1$.¶

For this purpose, we will use Gahuku-Gama intertribe relations dataset, as shown in figure 3 and first reported by Read in 1954 (3). The graphical representation of this network shows a clear partition into three communities with only two edges that disrupt the balance. We also note that the top, circular, community shows full structural balance ($Q_i = 0$), so we don't expect that any information will be able to pass into or out of this community.

For all of our simulations we will use parameter value of $\alpha = \frac{1}{2}$. In the case of the Gahuku-Gama data, though, this choice does not impact our results - the only type of edges that disrupt structural balance are positive ones between communities (all internal community edges are positive), and hence $N_i = 0 \; \forall \; i$.

Figures 4, 5 display how information spreads across the dataset for intermediary values of μ and ϵ. The behaviour we hoped for is clearly shown - all nodes in the two inter-connected communities are eventually activated. Further, we notice how the process initially spreads quickly across the community of the seed node and only later jumps across to the second community. At this point the entire second community is quickly activated.

Figures 6, 7 show the effect of the process when starting from an alternative seed node. 6 chooses a seed node in the second of the inter-connected communities and we see a reverse of the behaviour from 5 - the process spreads quickly initially in the seed community and later spreads to the neighbouring community. In 7, however, the process is confined to the secluded cluster, unable to escape, as our model specified and we predicted.

Figures 8, 9 show the effect that adjusting our parameter values. As noted in the previous section, we can reduce the model to a single degree of freedom with choice of timescale, so we only consider varying the ϵ parameter. In 8 we adjust so that ϵ and μ are of the same order of magnitude (or identical, in this case), and we see that the process spreads much quicker to the second community with edges between the two much more likely to be traversed. Conversely, with a much smaller choice for ϵ, as in 9, the process takes many more time-steps to spread to the second community.

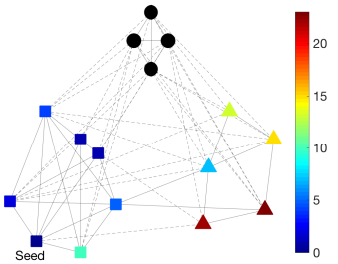

Fig. 8. Parameter Variation: $\mu = 0.5$, $\epsilon = 0.5$.¶

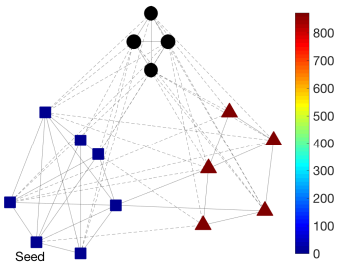

Fig. 9. Parameter Variation: $\mu = 0.5$, $\epsilon = 0.01$.¶

Discussion

Efficiency. It is important to note that the model can be implemented to run efficiently. The algorithm terminates in polynomial (quadratic) time and, crucially, this allows us to scale up to much larger datasets without excessive computational cost. Data on social networks often spans many thousands of nodes, see (12), and it is important to check that our model is able to deal with such networks.

Community Detection. During the course of this project we have devoted little time to the problem of dividing networks into communities and this will become a more significant issue for larger datasets. There are a significant number of papers, (9) (10) (11) for example, which provide a variety of methods for achieving this and we must pick the most appropriate method for each dataset individually. To complement the speed of our algorithm it is necessary to choose an efficient method and therefore Doreian's iterative method (8) is not recommended.

Parameter Values. Another topic that we have not covered in great detail is the choice of our model parameter values, α, μ, ϵ. The choice of α is personal and dependent on whether one wishes to place more weight onto negative edges within communities (higher values of α), or onto positive edges between different communities (lower values of α). Our argument of timescale rescaling allows us to eliminate one of the remaining parameters and without loss of generality we take $\mu = 0.5$. The value of ϵ is then highly dependent on the data and should be determined experimentally.

Further Research. The most obvious next step to this research is to test this model with a much larger dataset. There are, however, a number of interesting alternative applications to explore:

1. **Determining Community Structure:** One could reverse the problem and attempt to decompose a network's community structure using this model. If we are able to observe data on the spread of information through a network then we can test this against different model community configurations.

2. **Node Centrality:** Taking the average activation time of nodes across many simulations with different initialisations determines a measure of node centrality. Lower mean activation times signify a more prominent role in the spreading process. This notion of centrality is similar to geodesic centrality, counting the number of shortest paths through a node. An average over 100 simulations with a random seed is shown for the Gahuku-Gama data in figure 10.

3. **Analytical Properties:** This work has focused on numerical simulations and not analytical results with the probabilistic element we have incorporated presenting an additional challenge. At the very least, we could hope to derive results for expected jumping times between communities and such results would help to inform our parameter choices.

4. **Random Graphs:** This model hints at a different way of viewing random graphs in social networks - focusing on global community structure instead of finer connections. We could motivate a random graph formulation using this idea, parameterising community distribution and inter-community connections.

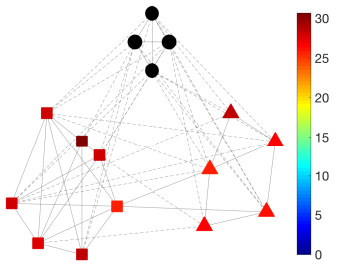

Fig. 10. Average activation time in Gahuku-Gama data over 100 simulations. $\mu = 0.5$, $\epsilon = 0.1$. All nodes in the two inter-connected communities are shown to be approximately equal under this version of centrality.

Conclusion

We have motivated a new way of thinking about dynamic processes on signed networks by considering their global structure and a community decomposition. Our results have focused on the Gahuku-Gama dataset to test this model against the properties we desire and serves simply as an introduction to the theory. There are a huge number of routes to pursue having established these basic properties, a sample of which are discussed above.

It is important to note that this model has been explicitly physically motivated by information passing through social networks. In analysing networks with other physical interpretations we must therefore be careful in checking the assumptions before applying the model. For example, the theory does not apply so readily to epidemiology, where positive and negative connections have less impact in the spread of disease. This example is a case where the process is egocentric and has minimal dependency on global network structure.

1. M. Ehsani, M. M. Sepehri *Balanced clusters and diffusion processes in signed networks.* Journal of Industrial and Systems Engineering. Vol 7, No. 1, pp 104-117. 2014.
2. M. G. Everett, Stephen P. Borgatti *Networks Containing Negative Ties.* Social Networks. Volume 38, pp. 111-120. July 2014.
3. K. E. Read *Cultures of the Central Highlands, New Guinea* Southwestern Journal of Anthropology. 1954.
4. M. E. J. Newman *Networks: An Introduction*
5. J. A. Davis *Clustering and structural balance in graphs* Human Relations. 1967.
6. D. Easley, J. Kleinberg *Networks, crowds and markets: reasoning about a highly connected world* CUP. 2010.
7. P. Doreian *Generalized Blockmodeling* CUP. 2005.
8. P. Doreian, A. Mrvar *Partitioning signed social networks* Social Networks. pp 1-11. 2009.
9. P. Esmailian, M. Jalili *Community Detection in Signed Networks: the Role of Negative Ties in Different Scales* Scientific Reports. 2015.
10. R. R. Xue, Y. H. Ma, W. Lei *Community Detecting in Signed Networks based on Modularity* Scientific Reports. 2015.
11. P. Anchuri, M. Magdon-Ismail *Communities and balance in signed networks: A spectral approach* 2012.
12. Stanford Large Network Dataset Collection (SNAP) *https://snap.stanford.edu/data/*

¶ Nodes coloured proportional to activation time. Black nodes are never activated.