

# SCIENTIFIC REPORTS



OPEN

## Community Detection in Signed Networks: the Role of Negative ties in Different Scales

Pouya Esmailian<sup>1</sup> & Mahdi Jalili<sup>2</sup>

Received: 23 December 2014

Accepted: 21 August 2015

Published: 23 September 2015

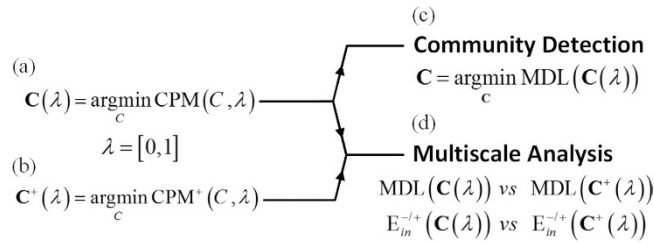
Extracting community structure of complex network systems has many applications from engineering to biology and social sciences. There exist many algorithms to discover community structure of networks. However, it has been significantly under-explored for networks with positive and negative links as compared to unsigned ones. Trying to fill this gap, we measured the quality of partitions by introducing a Map Equation for signed networks. It is based on the assumption that negative relations weaken positive flow from a node towards a community, and thus, external (internal) negative ties increase the probability of staying inside (escaping from) a community. We further extended the Constant Potts Model, providing a map spectrum for signed networks. Accordingly, a partition is selected through balancing between abridgment and expatriation of a signed network. Most importantly, multi-scale spectrum of signed networks revealed how informative are negative ties in different scales, and quantified the topological placement of negative ties between dense positive ones. Moreover, an inconsistency was found in the signed Modularity: as the number of negative ties increases, the density of positive ties is neglected more. These results shed lights on the community structure of signed networks.

During the last decade, there has been an ever-growing interest in community structure of real-world networks<sup>1,2</sup>. A community structure is observed on a network when relations are sparse, and there exists a mechanism driving density heterogeneity<sup>3,4</sup>. Indeed, grouping a network into modules of relatively high density provides novel insights into characteristics of a network<sup>5-9</sup>. However, the mainstream of this trend is mostly involved with only positive relations, where the links are absorptive, and thus “more links” means “more closeness”. In parallel with this trend, there has been some attempts towards community detection in networks with both positive and negative relations<sup>1,10,11</sup>. As a result, the objective is to partition a network into modules of relatively high density with as few internal negative ties as possible. Hence, this is different from those that are not bound to the density constraint, e.g., Correlation Clustering<sup>12</sup>.

A community is a group of tightly knitted nodes that are weakly connected to the rest of the network. This is not a *definition* per se, nonetheless it is the one upon which literature is mostly agreed, and can be further refined into specific definitions<sup>13</sup>. For signed graphs, the extension is straightforward by adding “as few internal negative ties as possible” criterion. Despite a vast literature on community detection, there has been few attempts towards signed networks<sup>1</sup>. Nonetheless, as the main step, Traag and Bruggeman generalized the modularity-alike objectives to signed graphs, as they have straightforward and intuitive signed counterparts<sup>14</sup>. However, the generalized form introduces a set of crucial parameters that cannot be trivially set. For the case of modularity, the objective function proposed by Gómez *et al.*<sup>15</sup> is as:

$$Q(G, C) = \alpha Q(G^+, C) - (1 - \alpha) Q(G^-, C), \quad (1)$$

<sup>1</sup>Department of Computer Engineering, Sharif University of Technology, Tehran, Iran. <sup>2</sup>School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia. Correspondence and requests for materials should be addressed to M.J. (email: mahdi.jalili@rmit.edu.au)



**Figure 1. The proposed method.** A spectrum of partitions  $C(\lambda)$  and  $C^+(\lambda)$  is created by applying the extended CPM on a signed graph (a), and its positive subgraph (b). The best community is the one with the lowest MDL (c). Also, by comparing  $C(\lambda)$  and  $C^+(\lambda)$  partitions based on MDL and the ratio of internal negative (positive) links, useful insights about the role and topological placement of negative ties are obtained (d).

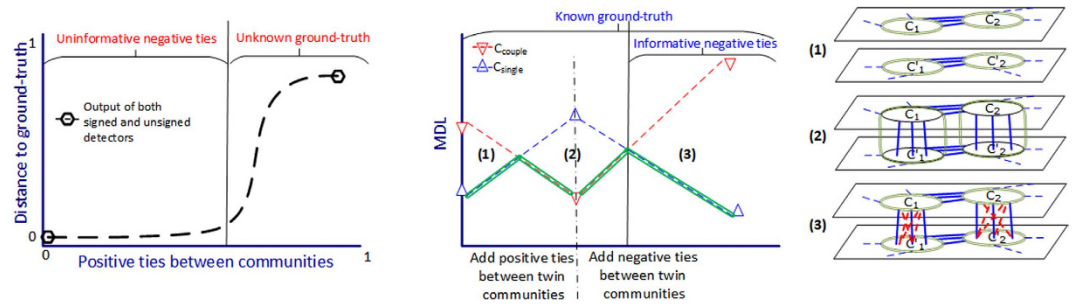
Symbol	Meaning
$E_a$	Number of directed links of induced graph $a$
$N_a$	Number of nodes in $a$
$\omega_a$	Sum of directed weights of induced graph $a$
$c_i$	Module of node $i$
$\omega_{ab}$	Sum of weights from $a$ to $b$
$\omega_a$	Sum of weights from $a$
$\omega_{i, in}$	Sum of weights from node $i$ to $c_i$
$\omega_{i, out}$	Sum of weights from node $i$ to all $c \neq c_i$
$ \omega $	Sum of absolute weights
$\omega_{..}^+$	Sum of positive weights
$\omega_{..}^-$	Absolute sum of negative weights
$a_{ab}$	$= \text{sign}(\omega_{ab})$
$P_{ab}$	$\omega_{ab}^+ / \omega_a^+$
$\delta(a, b)$	$= 1$ if $a = b$ ; $= 0$ otherwise

**Table 1. Basic notations.**  $a$  and  $b$  could be any set of nodes, but  $i$  is reserved for a single node. Notations defined for  $\omega$  could be inherited by others.

where  $G^+$  ( $G^-$ ) is the positive (negative) subgraph of  $G$ ,  $C$  is the partition, and  $0 \leq \alpha \leq 1$  is the relative importance of positive subgraph compared to the negative one, which has been set to  $E_G^+ / E_G^{15}$ . This objective function rewards (punishes) more positive (negative) density inside modules. In this work, we developed specialized algorithms for discovering community structure in signed networks. We first reformulated the Map Equation to measure the quality of partitions, known as Minimum Description Length (MDL). Next, we extended Constant Pots Model (CPM) to collect a spectrum of partitions from highly simplified to detailed ones, by sliding its parameter  $\lambda$  from zero to one (Fig. 1a,b). Based on these extensions, the community detection is carried out by minimizing MDL on  $\lambda$ -spectrum (Fig. 1c). Moreover, by comparing MDL and the ratio of internal negative (positive) links, the role and topological placement of negative ties can be quantified (Fig. 1d). As the experiments will show, the proposed method is highly reliable on both signed and unsigned networks, overcoming the resolution-limit and inconsistency of Modularity on signed networks. To evaluate the method, we proposed a novel benchmark in which negative ties can be informatively introduced. As a motivation, evaluating the effectiveness of signed detectors is possible only if the negative ties are informative and, to the best of our knowledge, this issue has not been addressed in previous works.

### Results

**Notation.** Throughout the paper, the expressions “link,” “edge,” “tie,” and “relation” are used interchangeably, unless we explicitly make a note. A graph  $G$  is determined by the triple  $(V, E, W)$  where  $V$  is the set of nodes,  $E$  is the set of edges defined by pair  $(v_i, v_j)$  of nodes  $((v_i, v_j) = (v_j, v_i)$  for undirected graphs), and  $W$  assigns a weight to each edge. The induced graph of node set  $a$  is defined by  $(V', E', W')$  where  $i) V' = a, ii) (v_i, v_j) \in E',$  if  $v_i, v_j \in a$  and  $(v_i, v_j) \in E,$  and  $iii) W'(v_i, v_j) = W(v_i, v_j)$ . A partition  $C$  assigns each node to a module (i.e.,  $c_i$  is the module of node  $i$ ). Other basic notations are listed in Table 1.

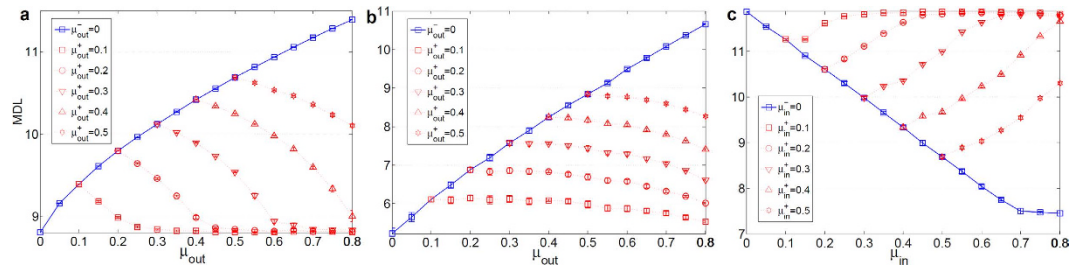


**Figure 2. Characteristics of the proposed benchmarks.**  $C_{single}$  ( $C_{couple}$ ) partition places each community of ground-truth (and its duplicate) in one module. (a) in signed LFR, for any arbitrary amount of external negative ties, by increasing positive ties between communities, the ground-truth is detectable by unsigned methods, as well as signed ones, until it fades out. (b) in coupled LFR, the duplicate (twin) communities are first intertwined with positive ties (phase 1 and 2) and then get separated by negative ties (phase 2 and 3). This leads to (A) a constantly known ground-truth (green double-line) which switches from  $C_{single}$  to  $C_{couple}$  and again back to  $C_{single}$ , and (B) informative negative ties in phase 3 where unsigned methods wrongly prefer  $C_{couple}$  to  $C_{single}$ . The transition points are illustrated using MDL measure (lower value corresponds to better partition).

**Tools to explore the community structure of signed networks.** In this work, we reformulated the well-known Map Equation  $L(G, C)$  for signed networks, which is the minimum expected code length that is required to address each step of a random-walker (also known as Minimum Description Length). The idea is that negative ties should be used in line with the “random walkers are more likely to be trapped inside a community” intuition. Therefore, a negative tie from node  $i$  of module  $c_i$  towards  $c'$  should decrease the probability of going from  $c_i$  to  $c'$ , and conversely, a negative tie towards the inside of  $c_i$  should increase the probability of escaping from  $c_i$ . Accordingly, Map Equation was reformulated to account for the negative ties (see Methods). By increasing the amount of external (internal) negative ties, it is expected that the value of MDL decreases (increases). Also, we extended the Constant Potts Model (CPM) to signed networks. CPM explicitly states that the absence of internal positive ties should be punished using a constant parameter  $\lambda$ . In particular, by sliding  $\lambda$  from 0 to 1 (or further), the minimization of CPM results in the extraction of smaller and denser modules<sup>16</sup>. Utilizing these well-established foundations, our method provided a map spectrum for signed networks, which not only revealed the best partition of a network but also provided useful information about negative ties on different scales of a network.

**Proposed benchmarks for signed networks.** In this manuscript, we extended the LFR (Lancichinetti-Fortunato-Radicchi<sup>17</sup>) benchmark to signed graphs. As the main characteristic, our extensions (the same as original LFR) provide scale-free distributions for positive (negative) degrees and community sizes. These benchmarks are denoted as signed and coupled LFR. The signed LFR simply replaces internal (or external) positive ties with negative ones. This extension was used to evaluate the effect of external (internal) negative ties on the extended Map Equation. However, in order to evaluate the power of signed detectors, a network must have two features: (1) a valid ground-truth for comparison, and (2) informative negative ties, which ignoring them leads to incorrect partitioning. These features do not simultaneously hold for the signed LFR, and thus, the coupled LFR was proposed. To show the failure of the signed LFR, the evaluation was started with uninformative negative ties, which can be ignored. Gradually, by placing positive ties between the communities, it was expected that at some point the negative ties become decisive, which means the unsigned detectors unlike signed ones must fail to detect the correct partition. However, before the total collapse of the ground-truth, this did not happen, signifying that the signed LFR has either uninformative negative ties or invalid ground-truth (Fig. 2a). Accordingly, the coupled LFR was introduced. Without going into details, it is composed of two identical LFR graphs, and the merging process occurs only between twin (duplicate) communities (Fig. 2b). As the schematic representation shows, the coupled LFR provides an interval (phase 3) with both informative negative ties and known ground-truth, suitable for evaluating the efficiency of signed detectors. The procedure of creating the benchmarks is as follows:

**Signed LFR.** LFR benchmark was extended by introducing external or internal negative ties regarding each node. For the case of external ones, there is an additional parameter  $\mu_{out}^-$ , where  $\mu_{out} = \mu_{out}^+ + \mu_{out}^-$ , that forces each node to have  $a_{i,out}^-/|a_i| \simeq \mu_{out}^-$ . For example, in a graph with  $\mu_{out} = 0.6$  and  $\mu_{out}^- = 0.25$ , regarding each node  $i$ , approximately 40% of links are positive and inside module  $c_i$ , 25% are negative and outside  $c_i$ , and 35% are positive and outside  $c_i$ . For the case of internal negatives, using parameter  $\mu_{in}^- = a_{i,in}^-/|a_i|$ , the procedure is the same as the external ones.



**Figure 3. SiMap on signed LFR.** (a) Network of two communities of size 500. We added the external positive ties until  $\mu_{out}^+$ , and the negative ties afterwards. Networks with  $\mu_{out}^+ = \mu_{out}^-$  had almost the same quality as the case with  $\mu_{out} = 0$ . (b) Network of many communities with  $N_G = 1000$ ; although  $\mu_{out}^- \geq \mu_{out}^+$ , the external negative ties may not cancel the positive flow from each node towards every community; therefore, dashed MDL curves dropped slowly, and never reached the level of  $\mu_{out} = 0$ . (c) Network of many communities with  $N_G = 1000$ ; the internal negative ties canceled positive ones almost at the same level as  $\mu_{in} = 0$ . Each value was averaged over 25 graph realizations.

**Coupled LFR.** This benchmark was built using two identical unsigned LFR graphs that were intertwined using parameters  $\mu_c$  and  $\mu_c^-$ , where  $0 \leq \mu_c < 1$  and  $\mu_c = \mu_c^+ + \mu_c^-$ . Thus, there are two layers of graphs where each node or community has a duplicate (or twin) in the other one. Keeping the layers unchanged, the benchmark forces node  $i$  to have  $a_{i,in}[\mu_c/(1 - \mu_c)]$  new links towards twin community  $c_i'$  (similar for  $\mu_c^-$ ). That is, by increasing  $\mu_c$ , twin communities become intertwined, and conversely, by increasing  $\mu_c^-$  they become separated. However, the nominal  $\mu_c$  may not be satisfied for some nodes, and thus, the empirical average was reported in the plots. It is worth mentioning that the “twin” notion is not a model of real-world networks, but an easy way of controlling the network structure to produce informative ties (Fig. 2b). In other words, by introducing the coupled LFR, we tried to ensure the effectiveness when negative ties are playing a decisive role in partitioning the network as well as to include some basic characteristics of real-world networks such as scale-free community sizes and node degrees.

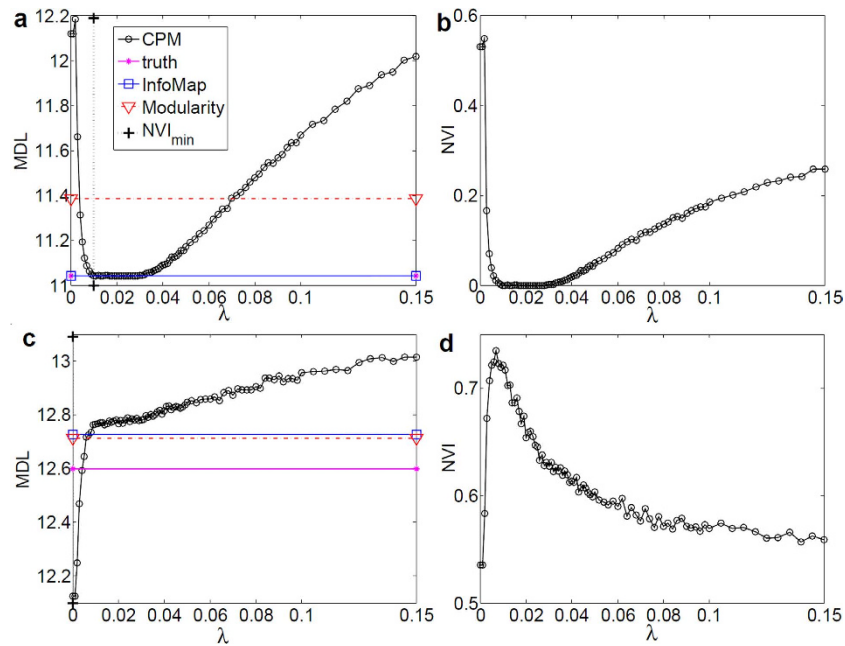
**Comparison of partitions.** The distance between two partitions  $C$  and  $C'$  was measured using Normalized Variation of Information (NVI). NVI is zero if the partitions are identical, and one if they are statistically independent, meaning no information is gained about  $C$  by knowing  $C'$  and vice versa; see the formulation in Methods.

The notations used for denoting general types of partitions are as follows:

- $C_{truth}$ : ground-truth partition of a graph for [un]signed LFR.
- $C_{Allin1}$ : all-in-one partition, which places all the nodes in one module.
- $C_{single}$ : places each community of coupled LFR in one module.
- $C_{couple}$ : places each community and its twin in one module.

**Evaluation of SiMap.** In order to investigate the effect of negative ties, SiMap was examined on signed LFR. In Fig. 3a,b, the internal structure of the communities was kept constant during the increase of mixing  $\mu_{out}$ , and similarly, the external links were kept constant during the increase of internal ties ( $\mu_{in}$ ) in Fig. 3c. As shown in Fig. 3a, for a network of two communities, when the mixing of only positive ties was increased, the value of MDL (solid curve) increased accordingly, which corresponds to the decrease of quality of communities. Next, we stopped adding positive ties at  $\mu_{out}^+$  and started adding negative ones afterwards, where  $\mu_{out} = \mu_{out}^- + \mu_{out}^+$ . The external negative ties are expected to cancel out the positive ones, and thus the quality of communities increases again (equally MDL decreases) almost to that of  $\mu_{out} = 0$ . This expectation was validated using different starting points (dashed curves in Fig. 3a). On the other hand, for a network of more than two communities, randomly-added external negative ties may not cancel the positive ties from each node towards every community. In other words, even if  $\mu_{out}^- \geq \mu_{out}^+$ , a node may have more positive ties toward a module than negative ones. Hence, as depicted in Fig. 3b, MDL dropped towards the level of  $\mu_{out} = 0$  with a slower slope and never reached that level. This is consistent with our formulation, stating that the external negative ties would cancel all the positive ones, if their weight towards every community is, at least, as much as positive ones, otherwise, MDL should be higher than that of  $\mu_{out} = 0$ . For the case of internal negative ties, a similar experiment was carried out. As shown in Fig. 3c, internal negative ties canceled the effect of positive ones, and consequently weakened the quality of communities almost to the situation where there had been no ties inside the communities ( $\mu_{in} = 0$ ).

In general, SiMap not only punishes the presence of internal negative ties, but also rewards the external negative ones. Note that the rewarding is module-wise, which means the mesoscopic topology of the



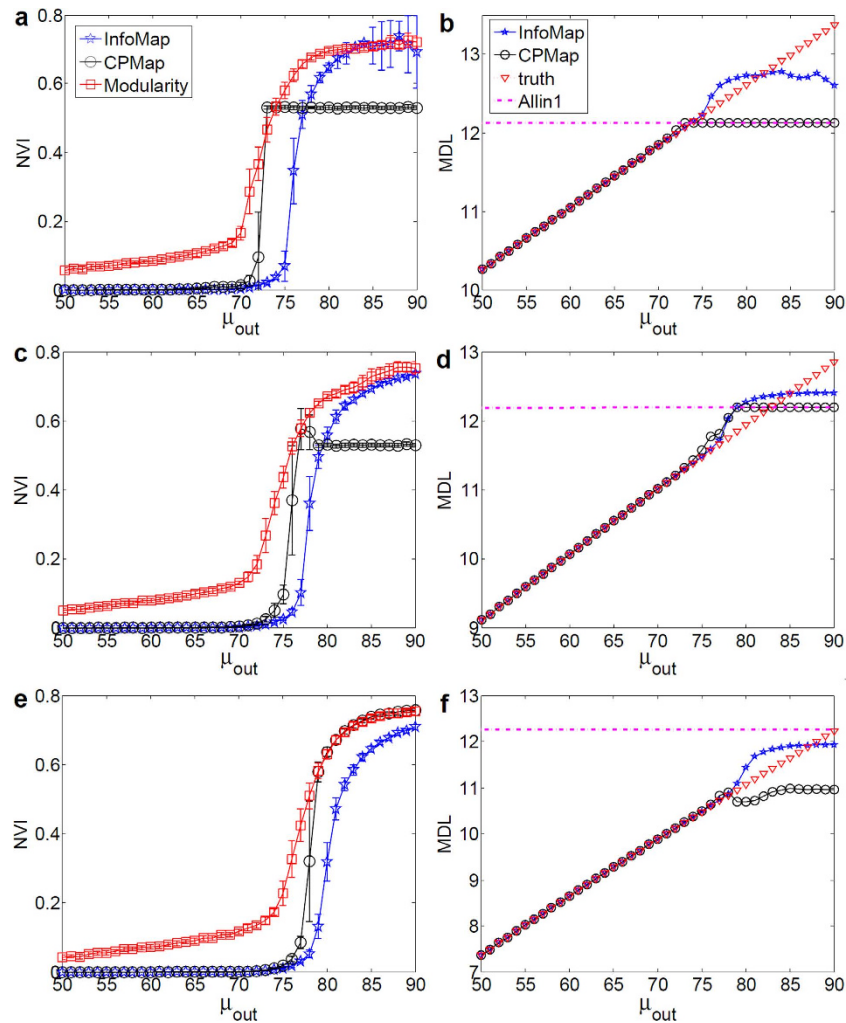
**Figure 4.** MDL spectrum of  $CPM(\lambda)$  on LFR ( $N_G = 5000$ ). (a,b) MDL and NVI for  $\mu_{out} = 0.6$ ; although  $\mu_{i,out} > \mu_{i,in}$  for every node, the community structure is still clear and detectable<sup>24</sup>. (c,d) MDL and NVI for  $\mu_{out} = 0.8$ , where the network has no significant community structure, and thus single-module has lower MDL than the case when the network is divided into sub-modules.

network determines the amount of the reward; e.g., if two modules have no positive ties in between, the inter-negative ones add no information, and therefore have no effect on MDL.

**Map spectrum of  $CPM(\lambda)$ .** In this experiment, the SiMap of signed  $CPM(\lambda)$  was plotted to illustrate its well-behaved curve with respect to the distance function  $NVI(C_{CPM}, C_{truth})$ . For comparison, the statistics of InfoMap (which ignores negative ties) and Modularity were also plotted. As depicted in Fig. 4a, although NVI is aware of the ground-truth partition, MDL curve behaved similarly to NVI, which first smoothly decreased, and then slowly increased. Additionally, the minimum of  $MDL(\lambda)$  coincided with the true community structure of the graph. Furthermore, InfoMap reached the same minimum level of MDL, meaning that the negative ties were not informative in this network. On the other hand, at  $\mu_{out} = 0.8$ ,  $MDL(\lambda)$  constantly increased from  $\lambda = 0$  to 1 (Fig. 4c). This indicates that the single-module is preferred to dividing the network into sub-modules, since it has no significant agglomeration of density at this mixing rate<sup>16</sup>.

**Performance of CPM on Signed LFR.** In this experiment, we compared InfoMap (which ignores the negative ties), CPM, and Modularity on signed LFR for  $\mu_{out}^- = 0$  (unsigned),  $\mu_{out}^- = 0.2$  and  $\mu_{out}^- = 0.4$ . In all cases, CPM performed better than the signed Modularity. As demonstrated in Fig. 5a,b, on unsigned LFR, CPM performed nearly as well as InfoMap in optimizing equation (15). This suggests that CPM provides a reach set of partitions. For  $\mu_{out} \geq 0.75$ , CPM opted for the single-module partition  $C_{Allin}$ , which had a lower MDL than both  $C_{truth}$  and the output of InfoMap. Nevertheless, according to Fig. 5a,d, there still remains a room for future improvements upon CPM to optimize SiMap. In Fig. 5c,d, by adding  $\mu_{out}^- = 0.2$  negative ties to each node, even InfoMap was still capable of detecting  $C_{truth}$  before  $\mu_{out} \simeq 0.75$ . After this threshold, all detectors failed to detect  $C_{truth}$ , since the community structure was not valid anymore due to severe rewirings<sup>16,18</sup>. In Fig. 5e,f, by adding more negative ties ( $\mu_{out}^- = 0.4$ ), although CPM reached a better MDL than all other partitions, the situation remained almost the same, that is signed LFR either has non-informative negative ties or invalid  $C_{truth}$ .

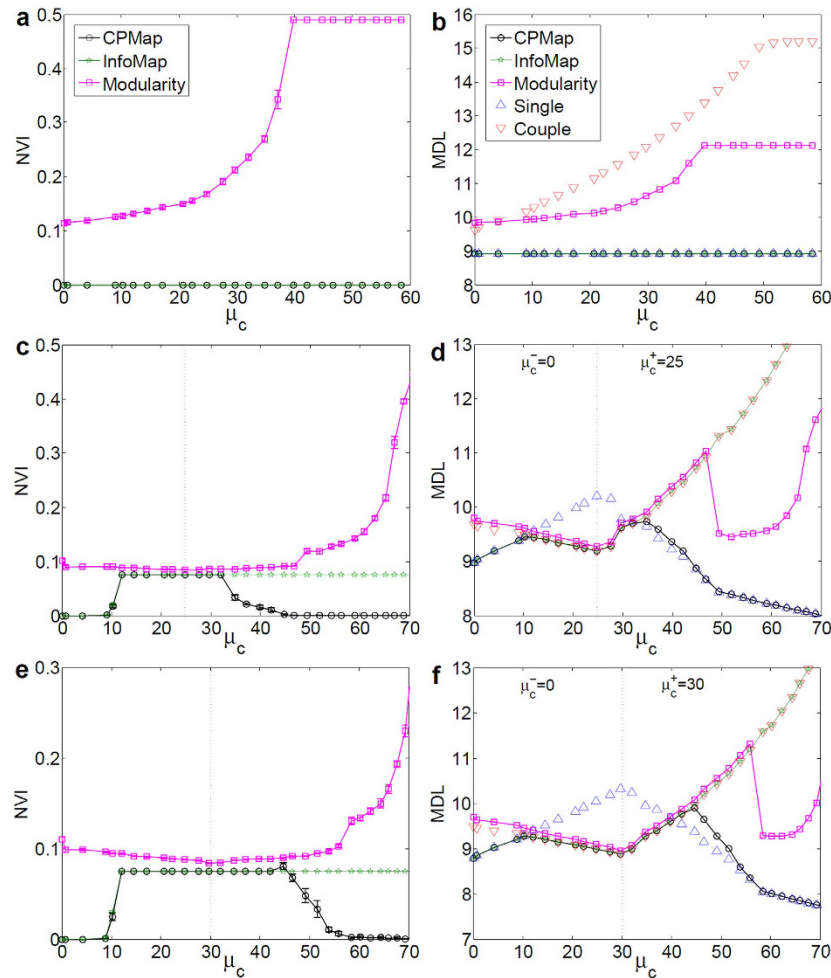
As a summary, although this benchmark may be first to come to mind, we showed it is not capable of appropriately challenging the signed detectors. The non-informativeness comes from the fact that flipping the external positive links to negative makes the community structure more clear, and thus InfoMap performs on signed LFR as accurate as the unsigned one. For this reason, the coupled LFR was introduced that gives us more leverage on the informativeness of negative ties while keeping the community structure valid.



**Figure 5. Evaluation of methods on signed LFR ( $\mu_{out}$  is plotted in percentage).** (a,b) NVI and MDL of the algorithms for unsigned LFR, (c,d) NVI and MDL for  $\mu_{out}^- = 0.2$ , and (e,f) for  $\mu_{out}^- = 0.4$ . In all cases, CPMMap and InfoMap had close performance and both outperformed Modularity. Before  $\mu_{out} \simeq 0.75$ , all detectors similarly performed on signed LFR as compared to the unsigned one. Roughly for  $\mu_{out} > 0.75$ , the community structure faded out and all detectors suddenly failed. Therefore, the signed LFR has either non-informative negative ties or invalid  $C_{truth}$ . Each value was averaged over 25 graph realizations with  $N_G = 5000$ .

**Performance of CPMMap on Coupled LFR.** In this experiment, using coupled LFR, we investigated the ability of CPMMap on utilizing the negative ties' information. To this end,  $\mu_{out} = 0.3$  was used for each layer to have connected yet well-separated communities, and only the connectivity of twin communities was manipulated. First, in Fig. 6a,b, two identical layers of graphs were gradually coupled only with negative ties ( $\mu_c = \mu_c^-$ ). As expected, the output of CPMMap and InfoMap constantly resembled  $C_{single}$ , since the negative ties added no competing information to the community structure of positive subgraph. However, the output of Modularity, surprisingly, changed with the increase of negative ties. In particular, by increasing  $\mu_c$ , previously detected modules were expanded to form larger ones. In other words, the number of negative ties indirectly weakened the sensitivity of Modularity to the density of positive ties, leading to larger and sparser modules.

Second, in Fig. 6c–f, the amount of positive ties  $\mu_c^+$  between twin communities was increased until the quality of  $C_{couple}$  surpassed that of  $C_{single}$ . At this point,  $C_{couple}$  was preferred by the detectors instead of  $C_{single}$ . Knowing that CPMMap and Modularity are partitioning the networks based on a criteria other than SiMap, this somehow validated the alignment of SiMap with the true quality of partitions. In the next phase, the negative ties  $\mu_c^-$  were added between twin communities to break them apart. As a result, the negative ties gradually became informative, since by ignoring them, InfoMap kept partitioning the graph exactly the same as  $C_{couple}$ . When  $C_{single}$  surpassed  $C_{couple}$ , CPMMap started to switch from  $C_{couple}$  to  $C_{single}$  accordingly. However, this switching occurred much later for the signed Modularity, meaning that

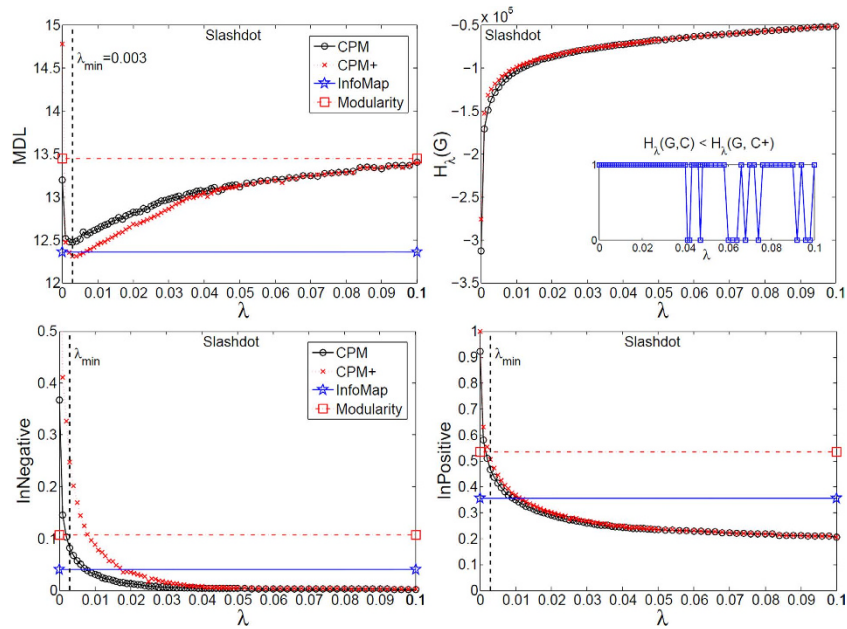


**Figure 6. Evaluation of methods on coupled LFR ( $\mu_c$  is plotted in percentage).** (a,b) Only negative ties were added between two layers.  $C_{single}$  constantly had the lowest MDL and was detected by InfoMap and CPMMap. However, the Modularity failed to detect  $C_{single}$  by merging its communities. (c,d) Positive ties were added between twin communities until  $\mu_c = 0.25$ , and negative ties were added afterwards. CPMMap constantly preferred the structure with the lowest MDL that is, in order,  $C_{single}$ ,  $C_{couple}$ , and again  $C_{single}$ . However, InfoMap wrongly preferred  $C_{couple}$  after  $\mu_c \simeq 0.1$ , signifying the usefulness of negative ties. (e,f) The experiment was repeated using more intertwined layers at  $\mu_c = 0.3$ . For all cases, Modularity reacted slowly to informative ties, and eventually failed to detect  $C_{single}$ . Each value was averaged over 25 graph realizations with  $N_{each\ layer} = 5000$ .

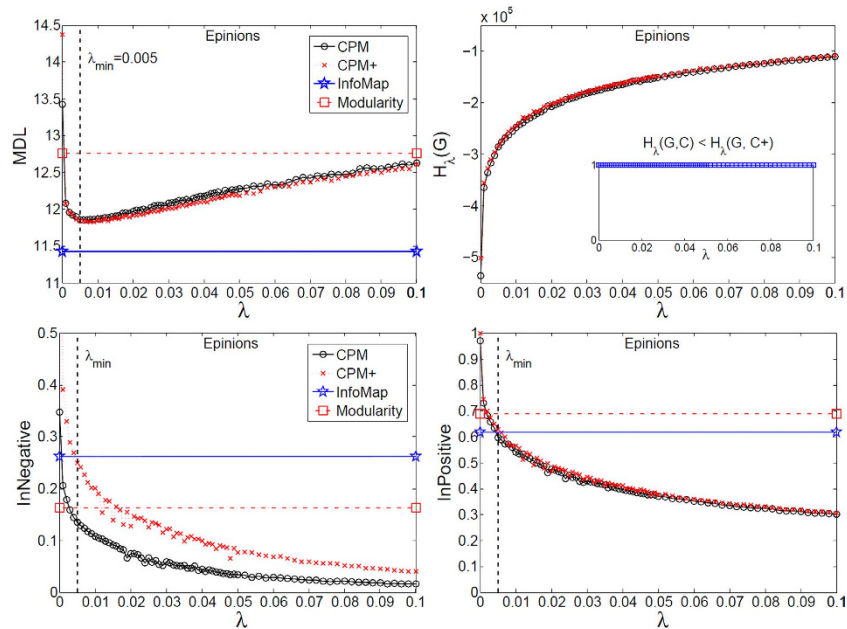
it is less sensitive to the informative negative ties. Also, it never opted for  $C_{single}$  due to an inherent inconsistency. Nevertheless, there is still a room for improvement upon signed CPM for optimizing SiMap, which is evident from  $0.3 < \mu_c < 0.45$  in Fig. 6c,d and  $0.45 < \mu_c < 0.55$  in Fig. 6e,f, where  $MDL_{single}$  is better than  $MDL_{CPM}$ .

**Online Social Media Networks.** Using the proposed tools, we investigated the mesoscopic structure of three well-known real signed networks: Slashdot, Epinions, and WikiElections. To this end, we optimized equation (17) for the whole spectrum of  $\lambda = [0, 1]$ , and further analyzed the corresponding partitions at each scale (only the informative intervals are plotted). In particular, the main information comes from comparing the spectrum of signed CPM to that of CPM+, which is only applied on the positive subgraph, to find the role of negative ties in different scales of the network. Also, this spectrum of partitions were compared with the output of InfoMap and signed Modularity.

As depicted in Figs 7, 8 and 9, MDL curve was V-shaped for all three networks with a minimum at  $\lambda_{min}$ , which signifies that *i*) the networks have community structure, and *ii*) the best map of each network is made up of modules with density  $\simeq \lambda_{min}$  or higher that are mutually connected with the same density or lower. In particular, WikiElections had considerably denser modules than both Slashdot and Epinions, which is consistent with its relatively higher density of positive ties (see Table 2). Also, regarding the



**Figure 7. Mesoscopic spectrum of Slashdot dataset.** MDL,  $H_\lambda(G, C)$ , and the ratio of internal negative (positive) ties are plotted for the output of CPM( $\lambda$ ), CPM+( $\lambda$ ), InfoMap, and signed Modularity.  $\lambda_{min}$  marks the best scale for CPM at which it has the lowest MDL. CPM+ is the application of CPM on positive subgraph. If the output of CPM has lower (better)  $H$  than that of CPM+, the value of inset will be 1, and 0 otherwise.

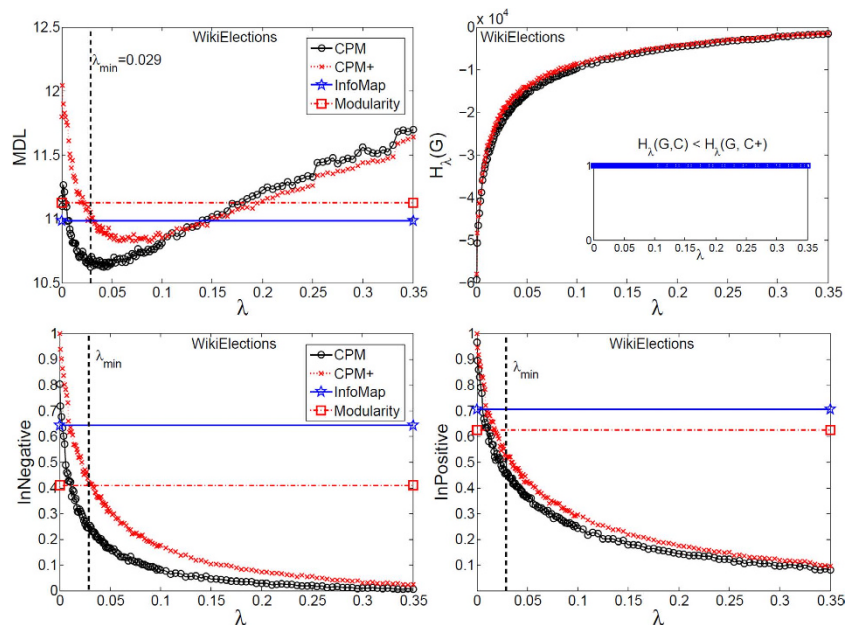


**Figure 8. Mesoscopic spectrum of Epinions dataset.** Designations are as Fig. 7.

internal negative (positive) ties, they continuously declined and were placed between the modules, as the interpretation of  $\lambda$  suggests.

**Slashdot.** According to Fig. 7, in all scales, the MDL curve of CPM+ was better than that of CPM. Also, MDL of InfoMap was better than MDL( $\lambda_{min}$ ) of CPM. Although the optimized value of equation (17) was slightly better for CPM than CPM+, which implies the constructive role of the negative ties in the optimization process, better MDL of both CPM+ and InfoMap signified that negative ties were not informative to achieve a higher quality mesoscopic structure as compared to the coupled LFR. Also, in





**Figure 9.** Mesoscopic spectrum of WikiElections dataset. Designations are as Fig. 7.

	$N_G$	$E_G/2$	$E_G^-/E_G$	$E_G^+/N_G^2$
Slashdot	75012	486537	0.2171	1.35E-04
Epinions	103157	668740	0.127	1.10E-04
WikiElections	6521	98907	0.2072	3.70E-02

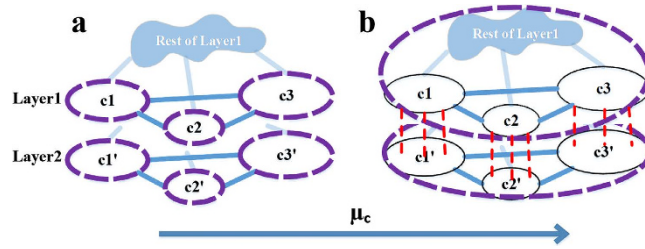
**Table 2.** Basic statistics of datasets after preprocessing.  $E_G/2$  is the number of edges,  $E_G^-/E_G$  is the ratio of negative edges, and  $E_G^+/N_G^2$  is the density of positive subgraph.

terms of the internal negative ties, CPM+ placed them between modules after  $\lambda \simeq 0.015$  without using their information, implying that almost all (95%) of negative ties were naturally placed between modules of density  $\simeq 0.015$  or higher that were mutually connected with the same density or lower. These findings are consistent with previous ones both on microscopic<sup>19</sup> and mesoscopic levels<sup>20</sup>. However, using the proposed tools, one captures a more quantitative picture of negative ties for the entire spectrum of the mesoscopic structure.

**Epinions.** According to Fig. 8, similar to Slashdot, CPM+ and InfoMap reached a better MDL than CPM, meaning that one could not find a better partition by taking negative ties into account. However, CPM+ could not exclude 95% of negative ties until  $\lambda \simeq 0.085$ . Therefore, the negative ties are placed between the modules of higher density with stronger interconnections compared to Slashdot. This means that the “negative ties lie between dense positive modules” pattern is apparent, yet, less salient than Slashdot.

**WikiElections.** According to Fig. 9, unlike Slashdot and Epinions, CPM had considerably better MDL than both CPM+ and InfoMap at the best scale  $\lambda_{min} = 0.0029$  and beyond until  $\lambda \simeq 0.1$ . This suggests that the information of negative ties is useful for WikiElections, and only vanishes when one zooms into the network to find the modules of density  $\simeq 0.1$  or higher. Accordingly, negative ties lose their informativeness for  $\lambda \geq 0.1$ . However, this threshold is well before the trivial case  $\lambda \simeq 1$ , where the objective is merely to find positive cliques. Additionally, CPM+ could not place 95% of negative ties between the modules until  $\lambda \simeq 0.25$ , meaning that the position of negative ties between dense positive ones is considerably less notable than Slashdot and Epinions. It is worth mentioning that Leskovec *et al.* also observed this different pattern of relations from local perspective, which have resulted in a weak cross-generalization of link prediction models, and also less accurate models for WikiElections<sup>21</sup>.

Indeed, these observations can be explained using the intuitive *information trade-off* of negative ties between local level (for sign prediction) and mesoscopic level (for community detection). That is to say, the more principled the negative ties between dense positive regions, the more accurate a link type can



**Figure 10. Schematic representation of inconsistency of Modularity on coupled LFR.** By adding negative ties between twin communities, eventually, Modularity placed each layer of  $C_{single}$  into one module (b).

be predicted given the information of its neighbors, and conversely, the less information they have for the task of community detection<sup>20</sup>.

**More on Signed Modularity.** Based on the results from coupled LFR, by increasing the negative ties, Modularity loses its sensitivity to the density of positive ties. Moreover, our experiments showed that even if two layers of coupled LFR are connected by only negative ties, again, the increase of negative ties leads to placing each layer in one module (Fig. 10b). Note that the coupled LFR is used to resemble two positive regions of a network (with heterogeneous densities internally) that are connected by negative ties. As an attempt to explain this observation, we considered a coupled LFR with fixed parameter  $\mu_{out}$ , which controls the mixing of each layer, and tunable parameter  $\mu_c$ , which adds [only] negative ties between two layers. Setting  $q = \mu_c(1 - \mu_{out}) / (1 - \mu_c)$ , the amount of negative ties relative to positive ones is

$$E_G^- = qE_G^+, \tag{2}$$

and the sum of positive/negative ties from each module is

$$E_{c_i}^+ = \frac{E_c^+}{1 - \mu_{out}}, \quad E_{c_i}^- = \frac{\mu_c}{1 - \mu_c} E_c^+. \tag{3}$$

Hence,  $Q$  can be rewritten in module terms as follows:

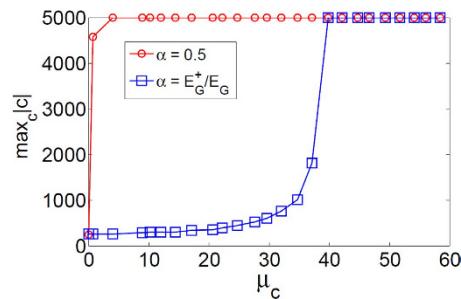
$$\begin{aligned} Q(G, C_{single}) &= \frac{E_G^+}{E_G} \sum_c \left( \frac{E_c^+}{E_G^+} - \left( \frac{E_{c_i}^+}{E_G^+} \right)^2 \right) - \frac{E_G^-}{E_G} \sum_c \left( \frac{E_c^-}{E_G^-} - \left( \frac{E_{c_i}^-}{E_G^-} \right)^2 \right) \\ &= \frac{1}{E_G} \sum_c \left( E_c^+ - E_c^- - (1 - q) \frac{E_c^{+2}}{\kappa E_G^+} \right), \end{aligned} \tag{4}$$

where  $\kappa = (1 - \mu_{out})^2$  is a constant value. According to the interpretation provided in Methods, the functionality of  $E_c^{+2}/E_G^+$ , similar to  $N_c^2$ , is to control the density of modules. Therefore, by increasing  $q$ , this sparsity-punishment term is attenuated, and consequently, modules try to gather more links ignoring the loss of density. That is to say, the *density* of each module gradually becomes less important, and conversely, the *number* of internal links becomes more important, which leads to larger and sparser modules consistent with our experiments (Fig. 11). In the same way, equalizing the importance of positive and negative ties ( $\alpha = 0.5$ ) leads to an even worse situation as:

$$Q(G, C_{single}) = \frac{1}{2E_G^+} \sum_c \left( E_c^+ - \frac{E_c^-}{q} \right), \tag{5}$$

implying that, for any  $q > 0$ , the objective is to have higher (lower) *number* of positive (negative) ties regardless of the density. This was revealed by our experiments which showed that each layer was placed into one module by a slight increase in  $\mu_c$  (Fig. 11). In fact, we argue that the reason for this failure is due to the implicit scale of Modularity that is similar for both positive and negative subgraphs. Setting  $\lambda^- = \lambda$  in equation (16), this becomes more clear as follows:

$$H(G, C) = \begin{cases} -\frac{E_G^+}{E_G} \sum_c (E_c^+ - qE_c^- - (1 - q)\lambda N_c^2) & \alpha = E_G^+/E_G \\ -\frac{1}{2} \sum_c (E_c^+ - E_c^-) & \alpha = 0.5 \end{cases}, \tag{6}$$



**Figure 11. Size of the largest module of Modularity on coupled LFR ( $\mu_c$  is plotted in percentage).** For  $\alpha = E_G^+/E_G$ , by increasing the number of negative ties between two identical LFR graphs (which have a clear community structure at  $\mu_{out} = 0.3$ ), modules were expanded until each layer of 5000 nodes was enclosed in one module. This happened after  $\mu_c \simeq 0.4$  corresponding to  $E_G^- \simeq \frac{1}{2}E_G^+$ . For  $\alpha = 0.5$ , complete expansion occurred right after a slight deviation from  $\mu_c = 0$ . The graphs were averaged over 25 realizations with  $N_{each\ layer} = 5000$ .

which basically leads the signed CPM to the same drawback as Modularity.

It can be concluded that when the number of negative ties increases, the sensitivity of Modularity to the *density* of positive ties decreases as the objective merely becomes grouping higher *number* of positive ties while excluding the negative ones, which is similar to Correlation Clustering. This inconsistency is resolved for CPM by setting  $\lambda^- = 0$  (see Methods).

## Discussion

In this work, we resolved the problem of community detection in networks with positive and negative edges. The proposed algorithm showed an excellent performance on novel synthetic benchmarks. Moreover, it provided a mesoscopic spectrum of signed networks, giving novel insights into the informativeness of negative ties as well as their topological placement between dense positive regions. Hence, one can attain a profound understanding about the structural relevance of positive and negative relations, and utilize that to justify the absorptive-repulsive behavior of the entities according to the context.

The proposed algorithm, CPM, showed a reasonable performance close to InfoMap on unsigned networks and non-informative signed networks, outperforming signed Modularity. Also, when the negative ties were informative, CPM performed excellent, extending the capabilities of InfoMap to signed networks. On the contrary, signed Modularity showed considerably weaker sensitivity to the presence of informative negative ties, as well as, growing inconsistency when the relative number of negative ties was increased. This inconsistency was further justified by the physical interpretation of the scale parameter  $\lambda$ , shedding new light on the general form of signed objectives.

Regarding the mesoscopic spectrum of real-world networks, we observed that negative ties in Slashdot and Epinions did not contribute to a better quality map than positive subgraph. However, they were informative for extracting the best map of WikiElections, where both CPM+ and InfoMap reached a similar MDL, yet, considerably worse than signed CPM. This usefulness lasted until we zoomed into the network to find the modules of density  $\simeq 0.1$  or higher. Moreover, the placement of negative ties between dense positive modules was more prominent in Slashdot and Epinions than WikiElections. However, this obscure pattern in WikiElections led to the extraction of more information from negative ties for community detection, consistent with the lower information extracted for the task of sign prediction<sup>21</sup>. Considering the nontrivial position of negative ties, if one wishes to detect modules of maximum density, i.e., positive cliques, negative ties obviously play no role in the detection task, and they are always placed between the modules. However, for Slashdot/Epinions/WikiElections, the majority of the negative ties were between modules of density  $\simeq 0.015/0.085/0.25$  or higher that were interconnected with the same density or lower, well before this trivial case. Therefore, we showed that it is expected to observe the “negative ties lie between dense positive ties” pattern in a nontrivial setting for real-world networks.

## Methods

**Tools to explore the community structure of signed networks.** We first introduce two objective functions used to determine the quality of communities: Map Equation<sup>22</sup> and Constant Potts Model<sup>23</sup>, which been previously used for unsigned graphs. We reformulated the Map Equation to signed networks (SiMap) by reweighting the walking patterns based on the mesoscopic information of negative ties. Also, we extended CPM to signed graphs, which remains unchanged when the same weight is used for both negative and positive terms. For the final algorithm, the only parameter of signed CPM,  $\lambda$ , is determined using SiMap.

**Map Equation.** Given a graph  $G$  and a partition  $C$ , Map Equation  $L(G, C)$  is the minimum expected code length that is required to address each step of a random-walker. Suppose a random-walker is going from node  $n$  to  $n'$ , this step is addressed as follows:

1. If the walker stays inside module  $c$ , a code is produced for  $n'$ .
2. If the walker goes from module  $c$  to  $c'$ , an exit-code for  $c$ , a code for  $c'$ , and finally a code for  $n'$  are produced sequentially.

Accordingly, there are two levels of coding. In the first level a code is assigned to each module, and in the second level each module receives a private coding for members and the action of exiting the module. Finally, using Shannon entropy, the theoretical minimum code length is achieved when the codes are assigned to entities based on their frequency of use. Consequently, the calculation of Map Equation is narrowed down to the relative frequency of visiting nodes and entering-exiting modules. Recent studies have shown that  $L$  is a very powerful criterion for detection of community structures, both experimentally<sup>24–26</sup> and theoretically<sup>27</sup>.

**Constant Potts Model (CPM).** To overcome the well-known resolution limit of modularity-like objectives, Traag *et al.*<sup>23</sup> suggested an objective function known as Constant Potts Model (CPM) as:

$$H(G, C) = -\sum_{i,j} (\omega_{ij} - \lambda) \delta(c_i, c_j), \quad (7)$$

where  $\lambda$  is a constant value. This equation can be rewritten in modular terms as:

$$H(G, C) = -\sum_c (\omega_c - \lambda N_c^2). \quad (8)$$

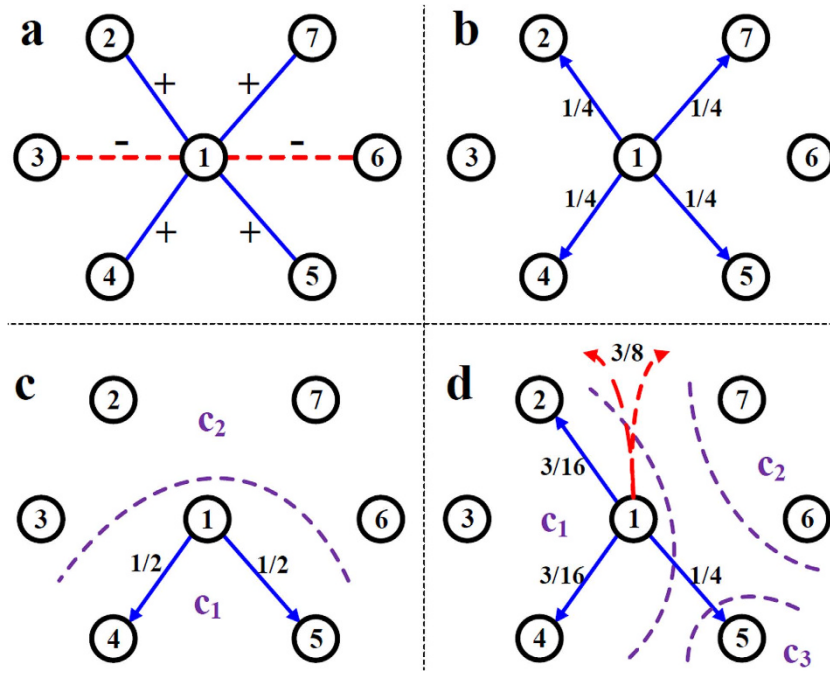
Theoretically, CPM has a clear interpretation based on  $\lambda$ <sup>23</sup>;  $H$  rewards module  $c$  with density  $\frac{\omega_c}{N_c^2}$  larger than  $\lambda$  and punishes  $c$  otherwise.  $H$  prefers modules  $r$  and  $s$  being separated if they have inter-density  $\frac{\omega_{rs} + \omega_{sr}}{2N_r N_s}$  smaller than  $\lambda$  and merges them otherwise. We used this interpretation to extend CPM to signed graphs and to analyze real-world networks. Although CPM has a simple formulation, it shows an outstanding performance on the state-of-the-art benchmarks if a proper  $\lambda$  is known a priori<sup>23</sup>. However, the burden of community detection is on parameter  $\lambda$ . One can get a wide range of partitions from all-in-one to each-in-one by sliding  $\lambda$  from 0 to 1 (and even further for weighted graphs). In other words, by increasing  $\lambda$ , we zoom into the network to see smaller, denser groups that are interconnected more densely. Consequently, the optimal value of  $\lambda$  is a fundamental key to the success of the method.

**Map Equation for signed networks (SiMap).** According to the proposed idea, the information of negative ties should affect the flowing pattern of a random-walker. As a result, given a graph  $G$  and a partition  $C$ , the weight (selection probability) of positive ties from node  $i$  of module  $c_i$  towards module  $c'$  is first decreased proportional to the negative ties from  $i$  towards  $c'$ , and then the remaining probability ( $p_{i,back}$ ) is channeled back to the internal links. Hence, if a random-walker arrives at node  $i$ , it is less likely to select the links toward module  $c'$  and more likely to step back inside  $c_i$ . After this, the weight of internal positive ties is decreased proportional to the internal negative ties, and finally, the remaining probability (negative teleport  $p_i^{tele}$ ), which has been subtracted from the internal positive ties, is uniformly split upon all the nodes in the network. As a summary, in the presence of external negative ties, a walker is less likely to leave  $c_i$ , and conversely, due to internal negative ties, it is more likely to escape from  $c_i$  by choosing a random node outside  $c_i$ .

Generally, the reweighting process is a heuristic choice. Nonetheless, one can simply make the following assumptions: *i*) if the weight of negative ties toward  $c'$  is at least the same as positive ties, the walker should not go to  $c'$ , and *ii*) if the same situation holds for the links toward the inside of  $c_i$ , the walker should not directly step back inside  $c_i$ . Based on these, we propose the following reweighting formulation:

$$p_{ij}^{new} = p_{ij}' + \frac{p_i^{tele}}{N_G}, \quad (9)$$

where reweighted (teleport-free) probabilities are



**Figure 12. Application of the reweighting procedure on node 1.** (a) Neighborhood of node 1 in a signed network. (b) The flow pattern of positive subgraph from node 1, which is used in the unsigned Map Equation. (c) Node 1 has the same amount of negative and positive ties towards  $c_2$ , thus, all the flow is channeled back into  $c_1$ . (d) All the flow towards  $c_2$  is channeled back into  $c_1$ , the flow towards  $c_3$  remains unchanged, and  $3/8$  of flow is uniformly spread over the network due to internal negative tie  $1 \rightarrow 3$ , which increases both the disorderedness of flow and the probability of escaping from  $c_1$ .

$$p'_{ij} = \begin{cases} 0 & \omega_{ij} < 0 \\ \max\left(0, 1 - \frac{\omega_{i,c_j}^-}{\omega_{i,c_j}^+}\right) p_{ij} & j \notin c_i, \omega_{ij} > 0, \\ \left(1 + \frac{p_{i,back}}{p_{i,in}}\right) \max\left(0, 1 - \frac{\omega_{i,in}^-}{\omega_{i,in}^+}\right) p_{ij} & j \in c_i, \omega_{ij} > 0 \end{cases} \quad (10)$$

and the negative teleportation from each node is

$$p_i^{tele} = \begin{cases} p_{i,in} + p_{i,back} - \left(1 + \frac{p_{i,back}}{p_{i,in}}\right) \max\left(0, 1 - \frac{\omega_{i,in}^-}{\omega_{i,in}^+}\right) p_{i,in} & p_{i,in} > 0, \\ p_{i,back} & o.w. \end{cases} \quad (11)$$

which depends on the backward flow calculated as

$$p_{i,back} = \sum_{p_{i,c} > 0, c \neq c_i} \min\left(1, \frac{\omega_{i,c}^-}{\omega_{i,c}^+}\right) p_{i,c}. \quad (12)$$

To be cleared, two examples of the procedure, which is applied on a sample node, are provided in Fig. 12. As a special case, if there is no negative tie, all transition probabilities remain unchanged.

Now, the probabilities of visiting nodes and entering (or exiting) modules need to be calculated. Note that a graph must be ergodic in order to have a stationary visiting distribution. The ergodicity is guaranteed by the use of teleportation that is being at node  $i$ , a random-walker either teleports to node  $j$  with probability  $\tau \nu_j$  where  $\sum_j \nu_j = 1$ , or selects link from node  $i$  to  $j$  with probability  $(1 - \tau) p_{ij}^{new28}$ . Moreover, in an ergodic graph the probability of entering or exiting a module is the same. The probability of visiting node  $i$  in the reweighted graph  $G'$  can be recursively calculated as:

$$\begin{aligned}
 p_i^{t+1} &= \tau \nu_i \sum_j p_j^t + (1 - \tau) \sum_j p_j^t p_{ji}^{new} \\
 &= \tau \nu_i + (1 - \tau) \sum_j p_j^t p'_{ji} + \frac{1 - \tau}{N_G} \sum_j p_j^t p_j^{tele}.
 \end{aligned}
 \tag{13}$$

Since the graph is ergodic, starting with an arbitrary distribution, e.g.,  $p_i^0 = 1/N_G$ , equation (13) converges to the true visiting probabilities [Empirically, distance  $\|P^{t+1} - P^t\|_2$  drops to  $10^{-15}$  after around 100 iterations on a graph with  $10^6$  links.]. Having  $p_i$  calculated for each node  $i$ , the exiting (or entering) probability of a module  $c$  is:

$$\begin{aligned}
 Q_c &= \sum_{i \in c} \left( \tau \left( 1 - \sum_{j \in c} \nu_j \right) + (1 - \tau) \sum_{j \notin c} p_{ij}^{new} \right) p_i \\
 &= \tau \left( 1 - \sum_{j \in c} \nu_j \right) \sum_{i \in c} p_i + (1 - \tau) \left( \frac{N_G - N_c}{N_G} \sum_{i \in c} p_i p_i^{tele} + \sum_{i \in c} p_i \sum_{j \notin c} p'_{ij} \right).
 \end{aligned}
 \tag{14}$$

As a special case, if there is no negative tie and the teleportation is uniform  $\nu_i = 1/N_G$ , equation (14) is the same as the one derived in<sup>28</sup>.

Having equations (13 and 14) as the main ingredients, the extended Map Equation (SiMap) is calculated as:

$$L(G, C) = Q \log Q - 2 \sum_c Q_c \log Q_c + \sum_c P_c \log P_c - \sum_i p_i \log p_i,
 \tag{15}$$

where  $Q = \sum_c Q_c$  is the probability of using a first-level code, and  $P_c = Q_c + \sum_{i \in c} p_i$  is the probability of using the second-level code of module  $c$ . Knowing that we encapsulated the information of negative ties in  $p_i$  and  $Q_c$ , equation (15) is the same as that of unsigned graphs<sup>28</sup>.

**Smart teleportation.** Lambiotte and Rosvall<sup>29</sup> showed that although teleportation probability  $\tau$  rectifies the instability of the visiting distribution, it may considerably bias the results. As a solution, they effectively resolved this biased behavior of Map Equation by: 1) setting  $\nu_i = \omega_i / \omega_G$ , 2) setting  $\tau = 0$  in equation (14), and 3) iterating equation (13) one extra step using  $\tau = 0$ . In this work, after the reweighting procedure, we did the same for calculation of equations (13 and 14).

**CPM for signed graphs.** According to the extension of modularity-alike functions to signed graphs<sup>14</sup>, CPM can be extended as:

$$H(G, C) = -\alpha \sum_c (\omega_c^+ - \lambda N_c^2) + (1 - \alpha) \sum_c (\omega_c^- - \lambda^- N_c^2),
 \tag{16}$$

which introduces a new parameter  $\lambda^-$ . In the same work, Traag and Bruggeman manually set non-zero values for parameters  $\lambda$  and  $\lambda^-$  to highlight their importance. Nevertheless, we argue that  $\lambda^-$  must be set to 0 for the case of CPM. The reason lies in the qualitative objective that is to detect “dense positive” and “negative-free” modules. In particular, “negative-free” condition can be restated as: *any* density of internal negative ties must be punished. However, according to the interpretation of  $\lambda$ , if density of the negative ties inside a module  $c$  is at most  $\lambda^-$ ,  $c$  receives an extra reward from equation (16). This implies that the internal negative ties wrongly increase the quality of  $c$  compared to the case when they are completely ignored. Therefore, by setting  $\lambda^- = 0$ , *any* amount of internal negative ties is punished. Also, if  $\alpha$  is set to 0.5 (equal contribution for positive and negative terms), both signed and unsigned CPM objectives will be the same, which only differ in multiplicative constant 0.5:

$$\begin{aligned}
 H(G, C) &= \frac{1}{2} \sum_c (\omega_c^+ - \lambda N_c^2) + \frac{1}{2} \sum_c \omega_c^- \\
 &= -\frac{1}{2} \sum_c (\omega_c^+ - \omega_c^- - \lambda N_c^2) \\
 &= -\frac{1}{2} \sum_c (\omega_c - \lambda N_c^2)
 \end{aligned}
 \tag{17}$$

Intuitively,  $\alpha$  is set to 0.5 to have a certain amount of positive ties being canceled out by the same amount of negative ones. Nonetheless, depending on the application, if intrinsic value of positive ties differ from negative ones, their weights must be set accordingly prior to applying the algorithm.

To optimize CPM, we use an improved version of Louvain method devised by Rosvall and Bergstrom, which is also the one utilized in InfoMap<sup>30</sup>. Louvian method first assigns a unique label to each node, then expands each label to those neighbors that maximally improve the objective value, and finally folds each module into a node and repeats the procedure until no further improvement is made<sup>31</sup>. The improved procedure first runs the Louvain algorithm, and then recursively refines both the nodes and modules to enhance the objective value further<sup>30</sup>. In our experiments, after 3 to 4 refinements the objective value was not considerably improved. Furthermore, the same procedure is used for the Modularity so as to eliminate the potentially biased comparisons due to different optimization procedures.

The main ingredient of Louvain method is the local-update formula<sup>31</sup>. Considering the unsigned CPM, when a set of nodes  $\kappa$  is moved from module  $c$  to  $c'$ , the local update becomes as<sup>23</sup>:

$$\Delta H = \omega_{c\kappa} + \omega_{\kappa c} - \omega_{c'\kappa} - \omega_{\kappa c'} + 2\lambda N_{\kappa}(N_{c'} - N_c), \quad (18)$$

where  $\kappa$  is considered in both  $c$  and  $c'$  for calculating  $N_c$  and  $N_{c'}$ . For the case of signed CPM, the extension is straightforward as:

$$\Delta H(G, C) = \alpha \Delta H(G^+, C) - (1 - \alpha) \Delta H(G^-, C), \quad (19)$$

reminding that  $\lambda$  is set to zero for the negative subgraph. Regarding equation (19), the positive and negative subgraphs are treated separately during the optimization process.

**Constant Potts Map (CPMap).** SiMap cannot be optimized via local methods of Louvain type, since a local change in a partition costs in the order of total links rather than local links. Indeed, the selection probability of positive ties must be updated according to the new position of negative ties, and thus the stationary distribution needs to be recalculated using equation (13). Nevertheless, SiMap still can be used to select among a set of candidate partitions. In particular, SiMap is used to find the best map of a network among the partitions provided by signed CPM. As Traag *et al.* showed, CPM provides a spectrum of maps that goes from highly simplified to highly detailed by sliding  $\lambda$  from 0 to 1<sup>16</sup>. Hence, as the main goal of Map Equation suggests<sup>22</sup>, one can use SiMap to select a map that balances between abridgment and expatiation, while constrained to the “negative free” condition.

Consequently, the proposed algorithm (CPMap) first feeds a set of  $\lambda$ s to equation (17), then minimizes the equation to get the corresponding partitions, and finally outputs the one with the lowest SiMap. Candidate  $\lambda$ s indeed can be chosen in a number of heuristic ways. However, in the experiments, we had the following observations: *i*) for a network with clear community structure, going from  $\lambda = 0$  to  $\lambda \approx 0.1$ , the MDL curve smoothly dropped, and it slowly rose by further increasing  $\lambda$ , and *ii*) for a network with no community structure, the MDL curve rose at the beginning of sliding  $\lambda$  away from 0, which means grouping the network as a whole was preferred to dividing it. Based on these observations, the following  $\lambda$ -selection is proposed:

1. Set  $x = x_{best} = 0$  and  $len = 0.1$ .
2. If  $len < L$ , output the partition of  $\lambda = x_{best}$ .
3. Consider  $N + 1$  equally spaced  $\lambda$ s in  $[x, x + len]$ . For newly added ones, optimize equation (17) and calculate MDL of corresponding partitions,
  - (a). If  $MDL(x)$  is the minimum, output the partition of  $\lambda = x$ .
  - (b). Else if  $MDL(x + len)$  is the minimum, set  $x = x_{best} = x + len$ , then go to (2).
  - (c). Else if  $MDL(x')$  is the minimum, set  $x_{best} = x'$ ,  $x = x' - \frac{len}{N}$  and  $len = \frac{2len}{N}$ , then go to (2).

In the experiments, we set  $N = 4$  and  $L = 0.005$ ; since the MDL curve had smooth changes, either increasing  $N$  or decreasing  $L$  did not considerably improve the results.

**Normalized Variation of Information (NVI).** This metric<sup>32</sup> measures the distance between two partitions, which is defined as:

$$\begin{aligned} NVI(C_1, C_2) &= \frac{1}{\log N_G} [H(C_1|C_2) + H(C_2|C_1)] \\ &= -\frac{1}{\log N_G} \left[ \sum_{c \in C_1, c' \in C_2} P(c, c') \log \frac{P(c, c')^2}{P(c)P(c')} \right], \end{aligned} \quad (20)$$

where  $H(\cdot, \cdot)$  is the conditional entropy, and

$$P(c, c') = \frac{|\{i | c_i^1 = c, c_i^2 = c'\}|}{N_G}, \quad (21)$$

where  $c_i^\kappa$  denotes the module of node  $i$  in partition  $\kappa$ . In particular, given one partition, NVI measures the remaining uncertainty about the other one. For example, if  $C_1 = C_2$ , given one partition, there is no uncertainty about the other one, and thus  $NVI(C_1, C_2) = 0$ . NVI is also closely related to Normalized Mutual Information (NMI)<sup>33</sup>; however, NVI is a metric<sup>32</sup> and has a clear interpretation<sup>34</sup>.

**Common parameters of benchmarks.** The parameters for artificial benchmarks were set as:  $\gamma = 1$ ,  $\beta = 2$ ,  $mean_i |a_i| = 20$ ,  $max_i |a_i| = 50$ ,  $min_c |c| = 20$ , and  $max_c |c| = 100$ . Moreover, regarding the other frequently used settings<sup>24</sup>, the conclusion drawn from each experiment remains the same.

**Online Social Networks and Data preprocessing.** We analyze three widely studied online signed social networks: *Slashdot*, *Epinions*, and *WikiElections*<sup>21</sup>. These datasets have been frequently used as benchmarks for studying signed social relations [All datasets are publicly available at <http://snap.stanford.edu>. For more detailed statistics refer to <http://konect.uni-koblenz.de/>]. They have special characteristics that make them suitable for the analysis of social relations. For example, all of the links either positive (for friendship or trust) or negative (for enmity or distrust) have been explicitly established by users, which means none of them has been inferred indirectly or asked from a person.

We performed some preprocessings on these datasets making them suitable for our purpose:

1. In order to get an undirected network, reciprocal links with inconsistent signs were omitted, and the remaining links were considered as undirected (inconsistent relations were 0.4%, 0.7%, and 1.6% of relations in *Slashdot*, *Epinions*, and *WikiElections*, respectively).
2. Only the largest connected component of each network was considered (99%, 90%, and 85% of nodes in *Slashdot*, *Epinions*, and *WikiElections*, respectively).
3. Nodes incident to zero positive edges were removed as they, trivially, belong to an isolated cluster.

Table 2 summarizes the properties of these networks after the above operations.

## References

1. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
2. Papadopoulos, S., Kompatsiaris, Y., Vakali, A. & Spyridonos, P. Community detection in social media. *Data Min. Knowl. Disc.* **24**, 515–54 (2012).
3. Newman, M. E. J. & Park, J. Why social networks are different from other types of networks. *Phys. Rev. E* **68**, 036122 (2003).
4. Backstrom, L., Huttenlocher, D., Kleinberg, J. & Lan, X. Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 44–54. Philadelphia, PA, USA: ACM. (2006 August).
5. Guimera, R. *et al.* Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, 065103 (2003).
6. Lancichinetti, A., Kivela, M., Saramaki, J. & Fortunato, S. Characterizing the community structure of complex networks. *PLoS ONE* **5**, e11976 (2010).
7. Onnela, J.-P. *et al.* Geographic constraints on social network groups. *PLoS ONE* **6**, e16939 (2011).
8. Grabowicz, P. A. *et al.* Social features of online networks: The strength of intermediary ties in online social media. *PLoS ONE* **7**, e29358 (2012).
9. Javari, A. & Jalili, M. Cluster-based collaborative filtering for sign prediction in social networks with positive and negative links. *ACM Trans. Intell. Syst. Technol.* **5**, 24 (2014).
10. Porter, M. A., Onnela, J.-P. & Mucha, P. J. Communities in networks. *Notices of the AMS* **56**, 1082–97 (2009).
11. Granell, C., Gómez, S. & Arenas, A. Mesoscopic analysis of networks: Applications to exploratory analysis and data clustering. *Chaos* **21**, 016102 (2011).
12. Bansal, N., Blum, A. & Chawla, S. Correlation clustering. *Mach. Learn.* **56**, 89–113 (2004).
13. Radicchi, F. *et al.* Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **101**, 2658–63 (2004).
14. Traag, V. A. & Bruggeman, J. Community detection in networks with positive and negative links. *Phys. Rev. E* **80**, 036115 (2009).
15. Gómez, S., Jensen, P. & Arenas, A. Analysis of community structure in networks of correlated data. *Phys. Rev. E* **80**, 016114 (2009).
16. Traag, V. A., Krings, G. & Dooren, P. V. Significant scales in community structure. *Sci. Rep.* **3**, 2930 (2013).
17. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**, 046110 (2008).
18. Leskovec, J., Huttenlocher, D. & Kleinberg, J. Signed networks in social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1361–70. Atlanta, Georgia, USA: ACM. (2010 April).
19. Esmailian, P., Abtahi, S. E. & Jalili, M. Mesoscopic analysis of online social networks: the role of negative ties. *Phys. Rev. E* **90**, 042817 (2014).
20. Leskovec, J., Huttenlocher, D. & Kleinberg, J. Predicting positive and negative links in online social networks. *Proceedings of the 19th International Conference on World Wide Web* 641–50. Raleigh, NC, USA: ACM. (2010 April).
21. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **105**, 1118–23 (2008).
22. Traag, V. A., Dooren, P. V. & Nesterov, Y. Narrow scope for resolution-limit free community detection. *Phys. Rev. E* **84**, 016114 (2011).
23. Lancichinetti, A. & Fortunato, S. Community detection algorithms: a comparative analysis. *Phys. Rev. E* **80**, 056117 (2009).
24. Aldecoa, R. & Marn, I. Closed benchmarks for network community structure characterization. *Phys. Rev. E* **85**, 026109 (2012).
25. Aldecoa, R. & Marn, I. Exploring the limits of community detection strategies in complex networks. *Sci. Rep.* **3**, 2216 (2013).
26. Orman, K. G., Labatut, V. & Cheri H. Comparative evaluation of community detection algorithms: a topological approach. *J. Stat. Mech.: Theory Exp.* P08001; doi: 10.1088/1742-5468/2012/08/P08001 (2012).
27. Kawamoto, T. & Rosvall, M. The map equation and the resolution limit in community. arXiv:1402.4385v1 (2014). Date of access: 04/05/2014.
28. Rosvall, M., Axelsson, D. & Bergstrom, C. T. The map equation. *E.P.J. S.T.* **178**, 13–23 (2009).



29. Lambiotte, R. & Rosvall, M. Ranking and clustering of nodes in networks with smart teleportation. *Phys. Rev. E* **85**, 056107 (2012).
30. Rosvall, M. & Bergstrom, C. T. Fast stochastic and recursive search algorithm. (2009). Available at: [www.tp.umu.se/rosvall/algorithm.pdf](http://www.tp.umu.se/rosvall/algorithm.pdf) (Accessed: 7th November 2013).
31. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* P10008; doi: 10.1088/1742-5468/2008/10/P10008 (2008).
32. Meilä, M. Comparing clusterings an information based distance. *J. Multivariate Anal.* **98**, 873–95 (2007).
33. Aldecoa, R. & Marn, I. Deciphering network community structure by surprise. *PLoS ONE* **6**, e24195 (2011).
34. Karrer, B., Levina, E. & Newman, M. E. J. Robustness of community structure in networks. *Phys. Rev. E* **77**, 046119 (2008).

### Author Contributions

P.E. and M.J. analyzed the map spectrum of signed networks. P.E. formulated/implemented the idea. P.E. and M.J. wrote the manuscript. M.J. reviewed the analysis.

### Additional Information

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Esmailian, P. and Jalili, M. Community Detection in Signed Networks: the Role of Negative ties in Different Scales. *Sci. Rep.* **5**, 14339; doi: 10.1038/srep14339 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>