# Evaluating balance on social networks from their simple cycles

PIERRE-LOUIS GISCARD*
*University of York, Department of Computer Science, UK*
*pierre-louis.giscard@york.ac.uk

PAUL ROCHET
*Université de Nantes, Laboratoire de Mathématiques Jean Leray, France*

AND

RICHARD C. WILSON
*University of York, Department of Computer Science, UK*

Signed networks have long been used to represent social relations of amity (+) and enmity (-) between individuals. Group of individuals who are cyclically connected are said to be balanced if the number of negative edges in the cycle is even and unbalanced otherwise. In its earliest and most natural formulation, the balance of a social network was thus defined from its simple cycles, cycles which do not visit any vertex more than once. Because of the inherent difficulty associated with finding such cycles on very large networks, social balance has since then been studied via other means. In this article we present the balance as measured from the simple cycles and primitive orbits of social networks. We specifically provide two measures of balance: the proportion $R_\ell$ of negative simple cycles of length $\ell$ for each $\ell \leqslant 20$ which generalises the triangle index, and a ratio $K_\ell$ which extends the relative signed clustering coefficient introduced by Kunegis. To do so, we use a Monte Carlo implementation of a novel exact formula for counting the simple cycles on any weighted directed graph. Our method is free from the double-counting problem affecting previous cycle-based approaches, does not require edge-reciprocity of the underlying network, provides a gray-scale measure of balance for each cycle length separately and is sufficiently tractable that it can be implemented on a standard desktop computer. We observe that social networks exhibit strong inter-edge correlations favouring balanced situations and we determine the corresponding correlation length $\xi$. For longer simple cycles, $R_\ell$ undergoes a sharp transition to values expected from an uncorrelated model. This transition is absent from synthetic random networks, strongly suggesting that it carries a sociological meaning warranting further research.

*Keywords*: Social networks; simple cycles; primitive orbits; and Monte-Carlo

2000 Math Subject Classification: 05C38, 90B10, 91D30, 91G60

## 1. Introduction

### 1.1 *Balance in networks*

Relations of amity and enmity between individuals are well represented by signed networks, where an edge is assigned a positive value if two individuals are acquainted and in good terms, and a negative one if they are instead enemies [9, 19, 21, 24, 37]. Such networks provide a natural setting to study inter-personal relationships and their correlations. For example, one could expect that people are friendly towards the friends of their friends, a situation that is said to be "balanced". More generally, on signed

networks, a group of individuals who are cyclically connected—i.e. forming a triangle, a square, a pentagon etc.—are said to be balanced if the number of negative edges in the cycle is even. Otherwise the cycle is said to be unbalanced. Sociologists have suggested that such negative cycles are the cause of tension and thus, that social networks should evolve into a state where balanced cycles are largely predominant [2, 3, 9, 24]. The question of whether this holds for real-social networks and if not, by how much this fails to be true, arose from these considerations in the 1940s [24].

Mathematically speaking, this sociological question translates into the following problem: on a signed network $G$, determine for all $\ell$ the percentage of negative simple cycles of length $\ell$. This problem remains largely unsolved owing to its natural formulation in terms of *simple cycles*—cycles which do not visit any vertex more than once. Unfortunately, enumerating all the simple cycles of a network *exactly* is computationally intractable. Indeed, this includes counting all the Hamiltonian cycles of the graph, a problem known to be #P-complete [8]. For this reason, we need to seek more efficient methods, which inevitably lead to some approximations. Two strategies are implemented in this work: i) approximate the balance of the network to within any desired accuracy by evaluating the balance on a large sample of subgraphs of the network; or ii) compute the balance exactly from objects which are not simple cycles, but should carry a similar information.

We successfully implemented the first strategy thanks to a novel exact formula for counting simple cycles on any (weighted directed) graph in conjunction with a Monte Carlo approach. This method is presented in Section 2. It effectively solves the mathematical problem enunciated earlier since the quality of the obtained approximation is controlled and can be improved at will. For the second strategy, we relied on the primitive orbits of the graph, cycles which contain no backtracking steps or tail, and are not the multiple of any other cycle. This is presented in Section 3. The results produced by both approaches on four social networks are discussed and compared in Section 4.

### 1.2 *Notation*

Throughout this article, we consider signed directed networks $G = (\mathscr{V}; \mathscr{E})$, of which undirected networks are a special case. The adjacency matrix of $G$ is denoted $\mathbf{A}_G$ or simply $\mathbf{A}$. Each edge of the network is weighted with a value +1 or -1 indicating a positive or negative interaction. A cycle is positive if the product of its edge values is positive, and otherwise it is negative. A cycle is simple if it does not visit any vertex more than once. The starting point of a simple cycle is irrelevant but its orientation is retained. For example, $v_0 v_1 v_2 v_0$ and $v_1 v_2 v_0 v_1$ represent the same triangle, which is however distinct from $v_0 v_2 v_1 v_0$. The number of positive and negative simple cycles of length $\ell$ on $G$ are designated by $N_\ell^+$ and $N_\ell^-$, respectively.

When discussing the balance of a network, we refer to the ratio $R_\ell$ of the number of negatively signed simple cycles of length $\ell$ to the total number of simple cycles of length $\ell$, i.e.

$$R_\ell := \frac{N_\ell^-}{N_\ell^- + N_\ell^+}, \quad \ell \geqslant 1.$$

In particular, $R_\ell = 0$ when the network is perfectly balanced for length $\ell$, while $R_\ell = 1$ indicates a totally unbalanced situation. To facilitate comparisons with existing results, we will also provide the ratio of negative to positive simple cycles $\mathscr{U}_\ell := N_\ell^- / N_\ell^+$ and the relative signed clustering coefficient $K_\ell := (N_\ell^+ - N_\ell^-)/(N_\ell^- + N_\ell^+)$ in Appendix C. In the next section, we will see that, for $\ell = 3$, the coefficients $R_\ell$ and $K_\ell$ are related to two measures of balance in social networks: the triangle index and the relative signed clustering coefficient (for triads), respectively.

### 1.3    *Existing approachs*

The study of balance in social networks has a long history and has been studied via a variety of approaches and we can thus only mention the latest works on the subject here. Of fundamental importance was the introduction of grayscale measures of balance, which quantify how balanced any given social network is [10, 31, 34]. Many such measures have been proposed, of which we shall briefly review a few important ones here.

The *signed clustering coefficient* and *algebraic conflict* evaluate the balance of a network locally—at the level of triangles—and globally, respectively [30, 31]. The former is defined as the number $T^+$ of positive triangles minus the number $T^-$ of negative ones divided by the total number of triplets of vertices connected by at least two edges. In other terms, the signed clustering coefficient takes into account the sign and the tendency to form small clusters (the triangles), putting both considerations on an equal footing. The relative signed clustering coefficient separates the two contributions, being the ratio of $T^+ - T^-$ to the total number of triangles, that is exactly the coefficient $K_3$ introduced earlier. The values of $K_\ell$ for $\ell \leqslant 20$ provided in this study can thus be seen as an extension of the relative signed clustering coefficient to larger clusters of up to 20 individuals. In general, the relative signed clustering coefficient $K_3$ is exactly accessible via the signed adjacency matrix, see Appendix C.

The algebraic conflict is defined on Harary's basic premise that "in a balanced network, all vertices can be divided into two groups such that all positive edges connect vertices within the same group, and all negative edges connect vertices of the two different groups" [18]. Kunegis [31], observed that this condition was fulfilled if and only if the smallest eigenvalue of the graph signed Laplacian is nil. This eigenvalue, called the algebraic conflict, thus carries information about balance and its distance to zero measures the failure of the network to be perfectly balanced.

Cycle-based approaches have also been devised, which evaluate various ratios involving the positive and negative cycles of a signed graph. These include the *degree of balance* and its weighted counterpart, arising either from weights attached to the graph edges or from additional length-dependent weights attached to the cycles, see e.g. [13]. The *triangle index* is the fraction of positive triangles on a graph to the total number of triangles. This index is $1 - R_3$ and can therefore be seen as a special case of the balance ratios studied in this work. Formally, obtaining the triangle index is a tractable task taking at most $O(N^3)$ time, see Appendix C for a formula giving $R_3$. This computational cost may nonetheless be too high on very large real-life networks. Instead, recent research has shown it could be reliably approximated by computing only the $k$ most dominant eigenvalues of the signed adjacency matrix of the network, thereby reducing the time needed to $O(N^2 k)$ [43]. Alternatively, the triangle index can be approximated via the Monte Carlo approach presented here.

In a 2011 study, Facchetti, Iacono, and Altafini [14] studied the global balance of social networks using the number of edges whose sign must be changed so that all the simple cycles—and hence all the cycles—be positive. This measure and its close variants are known as the *frustration index* [26, 33] and *line index of balance* [19, 20]. Although these were believed to be computationally intractable to obtain exactly [14], recent research shows otherwise [39], putting these quantities within reach on networks with a few thousands nodes. By studying the frustration index, Facchetti *et al.* found that social networks are indeed strongly balanced.Their conclusions were called into question in a recent work by Estrada and Benzi [13] who rather concluded that social networks are not so well balanced using a completely different cycle-based measure of balance, while still other studies employing the frustration index have instead confirmed the conclusion of Facchetti *et al.* [3]. As noted earlier, since counting all the simple

cycles of a large graph exactly is intractable, one may instead count objects which are not simple but carry a similar information when it comes to balance. In this vein, Estrada and Benzi [13] proposed the use of

$$D = \text{Tr} \exp(\mathbf{A}) = \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \text{Tr} \mathbf{A}^{\ell}, \tag{1.1}$$

as a method of counting the number of balanced and unbalanced *cycles*, also known as closed walks, in a network. They show that, by computing the ratio $K = D/D_+$ with $D_+ := \text{Tr} \exp(|\mathbf{A}|)$, called the *weighted degree of balance*, then the ratio of negative to positive cycles can be obtained as

$$\mathscr{U}^{\text{walks}} := \frac{1-K}{1+K}.$$

This is an extremely efficient method which simply requires the evaluation of the eigenvalues of the adjacency matrix. In practice, this could be expensive on very large networks, but it is sufficient to compute only a few of the largest eigenvalues of $\mathbf{A}$, as $D$ and $D_+$ are dominated by those [13].

Expression (1.1) counts all closed walks (weighted by a factor $1/\ell!$ for a walk of length $\ell$). Because of this, backtracking steps and multiple cycles are counted. For example, the non-simple cycle $v_0 v_1 v_2 v_0 v_1 v_2 v_0 = (v_0 v_1 v_2 v_0)^2$ is part of the sum. Such cycles are positive, and so they do not upset the balance of an already balanced network, but they do have an effect on the (global) balance ratio(s) for an unbalanced network. In other words, the cycle-sum embodied in (1.1) contains non-simple cycles from the order 2 onwards, effectively mixing the balance ratios $R_\ell$ at all lengths. In addition, the global signature of balance $\mathscr{U}_{\text{walks}}$ further mixes the contributions of the various cycles lengths. Figure 1 illustrates this issue in a triad. Whilst the network is completely unbalanced for triangles, we get $\mathscr{U}^{\text{walks}} = 0.19$, a number that is not easy to interpret.

These difficulties cannot be resolved easily using walks. In an attempt to better account for the length dependency of the balance, we define $D^\ell := \text{Tr} \mathbf{A}^\ell$, $D_+^\ell := \text{Tr} |\mathbf{A}|^\ell$ and $K_\ell^{\text{walks}} := D^\ell/D_+^\ell$, $\mathscr{U}_\ell^{\text{walks}} := (1 - K_\ell^{\text{walks}})/(1 + K_\ell^{\text{walks}})$ and

$$R_\ell^{\text{walks}} := \frac{D_+^\ell - D^\ell}{2 D_+^\ell}. \tag{1.2}$$

These quantities only take walks of length $\ell$ into account when calculating the balance. Yet, since short cycles and their multiples are typically much more abundant than long cycles, the values of $\mathscr{U}_\ell^{\text{walks}}$ and $R_\ell^{\text{walks}}$ are still largely dominated by the contributions from self-loops, backtracks and triangles. Consequently $R_\ell^{\text{walks}}$ will be depressed as compared to the true balance ratio $R_\ell$, that is, $R_\ell^{\text{walks}}$ overestimates the proportion of balanced cycles. We demonstrate this concretely in Section 4, Figures (2) and (3), where we compare $R_\ell^{\text{walks}}$ with $R_\ell$ calculated from the simple cycles on two social networks.

While one may empirically argue that long cycles are less relevant than short ones in real social networks [13, 47], it seems better to offer as detailed a mathematical analysis as possible before deciding this issue. For these reasons, we found it necessary to abandon the use of walks and rather recur either to the simple cycles themselves or to primitive orbits.

## 1.4 *Motivating our approach*

1.4.1 *On using simple cycles*   Earlier this year, Aref and Wilson [3] have produced a global study of existing measures of balance. In particular, they recommended the use of the frustration index as a
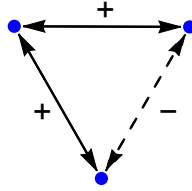
FIG. 1. An unbalanced network of three vertices. The dotted line represents a negative relationship. The exponential unbalance ratio is $\mathscr{U}_{\text{walks}} = 0.19$ while in fact $R_2 = 0$ and $R_3 = 1$.

measure of balance and raised several issues affecting cycle-based methods, which we briefly discuss below.

First, Aref and Wilson make an essential point concerning directed networks: many of the existing cycle-based studies do not take into account the fact that some edges are not reciprocated. It could be that an individual feels in good terms with someone who, in fact, feels negatively towards him. Or it could be that someone has passed a (positive or negative) judgment about someone else, whose attitude towards the first person is not reported in the data. This makes it difficult to interpret the results of approaches assuming edge-reciprocity, which should thus be, strictly speaking, limited to undirected networks. Second, Aref and Wilson write [3] "*cycle-based measures [...] are difficult to compute and do not provide a proper range of values, whether weighted or not.*" Finally, they note that such measures must be improved to avoid or diminish the influence of closed-walks with repeated edges or cycles— which they call cycle double-counting—as these are always balanced.

These observations are central to the present study, which addresses all of them. First, our approach is valid on directed signed networks, i.e. we do not assume nor need reciprocity of the edges[1]. Remark that an undirected graph is a special case of a directed one, where each directed edge $e_{ij}$ from $i$ to $j$ has a corresponding edge $e_{ji}$ of opposite orientation. It follows that any method for evaluating some indicator of balance that is valid on directed graphs gives rise to a valid approach for undirected graphs. Furthermore, a simple cycle on an undirected graph corresponds to two cycles with identical signs on the directed version of the same graph. Consequently, the balance ratios $R_\ell$ can always be calculated from the directed form of an undirected graph. Second, our approach offers a proper range of values, since $0 \leqslant R_\ell \leqslant 1$ for all $\ell$, which permits a gray-scale distinction between balanced or unbalanced situations on the one hand and the null hypothesis on the other. In the results, we will see that $R_\ell$ typically varies between 0 and 0.6, exploiting much of its theoretical range. Third, our method is sufficiently tractable that it can be implemented on a standard desktop computer (see below), even on large networks, thanks to a recent breakthrough in the combinatorics of cycles. Finally, our simple-cycle based method is rigorously free from the double-counting problem.

1.4.2 *On studying cycle-lengths separately* Existing approaches to social balance tend to fall into two categories, both of which provide useful but incomplete information:

▶ Methods based on small structures in social networks, in particular triads, are only sensitive to

---

[1]In other terms, the situations described by Aref and Wilson are exactly taken into account in our analysis. This motivated us to define the balance ratios $R_\ell$, for $\ell \geqslant 1$ rather than $\ell \geqslant 3$. And indeed, we will see below that $R_2 > 0$ in three of the four networks analysed, indicating that the lack of edge-reciprocity does occur in practice

local interactions in small groups of individuals. For this reason they are unable to detect long range correlations in relationships as well as characterise the dynamics of larger groups of individuals.

▶ Global measures of balance such as the degree of balance and algebraic conflict, are sensitive to social tensions occurring on all scales (that is group-sizes) but do not yield much information on where those tensions might be, as well as on the groups of individuals typically involved in balanced or unbalanced relations.

The approach recently taken by Facchetti and coworkers [14] is a step towards bridging this divide, as it provides both global and local information on the network balance. Yet, Estrada and Benzi have observed that the frustration index seems to be mostly determined by triads, leaving the influence of larger groups in the final result in doubt. In this work we have taken another route by extending two measures that have been developed for triads, the relative signed clustering coefficient and triangle index, to longer cycles. These longer cycles are sensitive to long range correlations between edge signs, and, by studying each length separately, provide a wealth of information on where social tension occurs by disentangling the influence of the various cycle-lengths on the overall balance.

## 2. Balance from simple cycles

### 2.1 *Core combinatorial result*

One possible strategy to estimate the balance ratios $R_\ell$ consists of approximating them from a large sample of subgraphs of the network under study. The main novelty permitting this straightforward approach in practice is a recently developed mathematical formula for counting simple cycles of any length on weighted directed graphs. Rather than sampling the simple cycles directly, the formula allows for a rapid and exact evaluation of the balance ratios on subgraphs of the original network. We show below that this strategy is much better from a computational standpoint than sampling the simple cycles themselves.

2.1.1 *Formula for counting simple cycles* Let $P(z)$ be the ordinary generating function of the simple cycles of any weighted directed graph $G$, that is

$$P(z) := \sum_{c:\text{ simple cycle}} w(c)\, z^{\ell(c)},$$

where $w(c)$ is the weight of $c$, that is the product of the weights of its edges, $z$ is a formal variable and $\ell(c)$ is the length of $c$. By exploiting algebraic structures associated with walks on graphs, $P(z)$ can be shown to be

$$P(z) = \int \frac{1}{z} \sum_{\substack{H \prec G \\ H \text{ connected}}} \mathrm{Tr}\left( (z\mathbf{A}_H)^{|H|} (\mathbf{I} - z\mathbf{A}_H)^{|N(H)|} \right) dz. \tag{2.1}$$

In this expression $H$ is a *connected*[2] induced subgraph of $G$, $\mathbf{A}_H$ its adjacency matrix, $|H|$ the number of vertices in $H$ and $|N(H)|$ the number of neighbours of $H$ in $G$. A neighbour of $H$ in $G$ is a vertex $v$ of $G$ which is not in $H$ and such that there exists at least one edge, possibly directed, from $v$ to a vertex of $H$ or from a vertex of $H$ to $v$. The result of Eq. (2.1), as well as further exact formulas for $P(z)$, is presented in [16] and shall not be proven here.

---

[2]If $G$ is directed, then the subgraphs $H$ should be *weakly connected* induced subgraphs of $G$. Recall that a digraph is said to be weakly connected if replacing all its directed edges by undirected edges produces a connected undirected graph.

2.1.2 *Computational cost* Let $n_H(G)$ be the number of connected induced subgraphs of a graph $G$ and let $n_c(G)$ be the total number of simple cycles on $G$. The main advantage of Eq. (2.1) is that, if we use a direct algorithm for finding all the connected induced subgraphs of a graph, for example using reverse-search, the time complexity is $O(N + |E| + N^2 n_H(G))$, with $N$ and $|E|$ the number of vertices and of edges of the graph, respectively [4, 11, 45]. In comparison a direct search of the simple cycles themselves must have a time complexity scaling with $n_c(G)$. The best general purpose algorithm for finding such cycles, Johnson's 1975 landmark algorithm [29, 35], has time complexity $O((N + |E|)(n_c(G) + 1))$, and on undirected graphs there exists an algorithm achieving the optimal $O(N(n_c(G) + 1))$ time complexity [6]. In spite of the fact that these algorithms can find the simple cycles, they are also commonly used when counting them is sufficient. In this situation however, Eq. (2.1) with its time complexity scaling with $n_H(G)$ represents a substantial speed-up since in general $n_H(G) \ll n_c(G)$. This is best demonstrated on the complete graph $K_N$ on $N$ vertices—which is the worst case scenario. On $K_N$ we have $n_c(G) \simeq e \times (N-1)!$ where $e \simeq 2.71828$ is the base of the natural logarithm, while $n_H(G) = 2^N$ is "only" exponential[3] Furthermore, in practice, the computational cost is much smaller as real networks are typically sparse.

Besides Johnson's algorithm, the most efficient *general purpose*[4] algorithm for counting simple cycles is that of Bax and Franklin [5]. This algorithm evaluates a formula counting the simple cycles that involves a sum over all induced subgraphs of the graph, as opposed to only the *connected* ones appearing in Eq. (2.1). On sparse graphs, there are far more induced subgraphs than connected induced subgraphs and evaluating Eq. (2.1) yields a significant speed-up as compared to Bax and Franklin's algorithm.

Most importantly for applications, Eq. (2.1) is well suited to truncations: only those connected induced subgraphs $H$ of $G$ for which $|H| \leqslant \ell \leqslant |H| + |N(H)|$ can possibly contribute to the coefficient of $z^\ell$ in $P(z)$. This means that if one is interested in the first $\ell$ terms of $P(z)$—that is in the simple cycles of length up to $\ell$—it suffices to consider those connected induced subgraphs of $G$ with $|H| \leqslant \ell$. Using reverse search, this has time complexity $O(N + |E| + \ell^2 n_{H:|H|\leqslant\ell}(G)) \leqslant O(N^2 + \ell^2 n_{H:|H|\leqslant\ell}(G))$ [11, 45]. Since furthermore only the small adjacency matrices $\mathbf{A}_H$ enter Eq. (2.1), each term of the equation costs $O(|H|^3) \leqslant O(\ell^3)$ to evaluate. Thus, getting the first $\ell$ terms of $P(z)$ from Eq. (2.1) has time complexity bounded by $O(\ell^3 N^\ell)$ in the worst case scenario of the complete graph, and *far less* on sparse graphs. Instead, the computational cost on sparse graphs can be bounded as follows: let $\Delta$ be the maximum vertex-degree on the network. Then the number of connected induced subgraphs on at most $\ell$ vertices is bounded by $O(N\Delta^\ell/((\Delta-1)\ell^2))$ [45]. In addition, the time complexity of reverse search in this case is $O(N(\Delta+1) + N\Delta^\ell)$ [4, 11, 45]. Consequently Eq. (2.1) can be shown to produce $P(z)$ exactly up to order $\ell$ in $O(N\Delta^\ell/\ell)$ time.[5] For a more detailed analysis of the algorithmic implementation of Eq. (2.1) together with comparison with other algorithms for the same task, see [38].

To give concrete examples, with an Intel Core i7-4790 CPU @ 3.60 GHz desktop computer, evaluating Eq. (2.1) on the complete graph on 15 vertices took on average $\sim 0.7$ sec, yielding $255,323,504,932 \simeq$

---

[3] One should keep in mind that since $P(z)$ determines the existence and number of Hamiltonian cycles on $G$, under the exponential time hypothesis [28] this exponential cost is, in principle, the best possible.

[4] In particular, the remarkable algorithm of Alon, Yuster and Zwick [1], which is extremely efficient for counting simple cycles, is not a general purpose one. Rather, it is limited to undirected graphs and cycle lengths of at most 7.

[5] The properties of Eq. (2.1) contrast it with other analytical formulas for counting simple cycles of small lengths [1, 22, 36]. First, these formulas work only on undirected graphs. Second, these formulas comprise a large number of terms, e.g. 160 terms for simple cycles of length 10, all of which involve the adjacency matrix of the full graph.

$2.5 \times 10^{11}$ for the total number of simple cycles, which we verify analytically to be exact. This went up to $\sim 25\,\mathrm{sec}$ for the complete graph on 20 vertices where $349,096,664,728,623,336 \simeq 3.5 \times 10^{17}$ simple cycles were counted, a number that is, once again, exact. In both cases about half of the computer time was spent looking for the connected induced subgraphs and the other half implementing Eq. (2.1). For the real-world networks analysed below, a randomly chosen induced subgraph on 30 vertices is typically analysed in $0.001 - 0.3\,\mathrm{seconds}$ on the same computer, depending on its sparsity.

## 2.2  *Monte Carlo implementation*

When the size of the network to study is large—what "large" means here strongly depends on the graph sparsity—an exact calculation of the desired terms of $P(z)$ from Eq. (2.1) becomes intractable and the core combinatorial result must be supplemented by a Monte Carlo approach.

The reader may have noticed upon close inspection of Eq. (2.1) that $P(z)$ results from subtle cancellations between the contributions of the various connected induced subgraphs $H$ of $G$. For this reason, Eq. (2.1) is not directly amenable to a Monte Carlo method which would consist of randomly selecting a sample of connected induced subgraphs $H$ of the whole network $G$ and estimating $P(z)$ from this sample. Eq. (2.1) can however be evaluated very quickly on graphs of "reasonable" size—once again this depends on the sparsity of the underlying graph and the available computational resources.

Our strategy is therefore to sample $N$ induced subgraphs of the network under study and to calculate the balance ratios $R_\ell$ *exactly* up to the desired length $\ell$ on each of these samples via Eq. (2.1). The average value of all the $R_\ell$ then converges to that of the whole network as $N$ grows. The quality of this approximation is appraised by repeating the whole procedure $N'$ times and extracting the standard deviation on the averaged $R_\ell$. If this deviation is too large, the number $N$ of samples is increased and the standard deviation is reevaluated. Once the deviation is below the desired accuracy, the method is deemed to have converged. We also systematically tested the method against bias by comparing it with exact results whenever available, see Appendix A.

Firstly, we decided to limit our study to simple cycles of length up to 20. Longer cycles are both rare and largely irrelevant since, as will see, at such lengths $\ell > 20$ all the social networks follow the null hypothesis. There remains three important parameters controlling the Monte Carlo approach which we can choose using a qualitative analysis:

1) Number of samples $N$: it is the total number of subgraphs on which the simple cycles are counted. The larger this number, the more simple cycles are found and the smaller the variance $\mathrm{var}(R_\ell)$ of the $R_\ell$ ratios, that is the better the results. Following well established principles, we expect $\mathrm{var}(R_\ell) \propto 1/\sqrt{|\mathrm{Cycle}_\ell|}$, $|\mathrm{Cycle}_\ell|$ being the total number of simple cycles of length $\ell$ found in all samples.

2) Size $N_{\mathrm{vertices}}$ of each sample: it is the number of vertices in each sampled subgraph. Since the longest simple cycle on a graph on $N_{\mathrm{vertices}}$ has length $\ell \leqslant N_{\mathrm{vertices}}$, we must have $N_{\mathrm{vertices}} \geqslant 20$. Letting $f_\ell(N_{\mathrm{vertices}})$ be the average number of simple cycles of length $\ell$ found on all sampled subgraphs, we have $|\mathrm{Cycle}_\ell| = N_{\mathrm{sample}} f_\ell(N_{\mathrm{vertices}})$. In addition, we can expect $f_\ell$ to be monotonous increasing function of $N_{\mathrm{vertices}}$, in particular $f_\ell(N_{\mathrm{vertices}})$ is at most $O(N_{\mathrm{vertices}}^\ell)$ on complete graphs.

3) Time $\tau_\ell(N_{\mathrm{vertices}})$ spent per sample: we have shown in the preceding section that this time obeys $\tau_\ell(N_{\mathrm{vertices}}) = O\left(N_{\mathrm{vertices}} \Delta^\ell / \ell\right)$, where $\Delta$ is the average maximum degree of any vertex in the

sampled subgraph. On average, $\Delta \leqslant N_{\text{vertices}}$ is expected to be a monotonous increasing function of $N_{\text{vertices}}$.

Taking these three quantities into account, we end up with the following trade-off between accuracy of the results and time taken to reach them

$$\text{var}(R_\ell) \propto \frac{1}{\sqrt{N f_\ell(N_{\text{vertices}})}} \iff \text{Total time} = O\left(N \frac{N_{\text{vertices}}^{\ell+1}}{\ell}\right),$$

from which we deduce that $N$ should be large while $N_{\text{vertices}}$ should be as small as possible. Indeed, the time taken to count simple cycles of length $\ell$ can scale as badly as $N_{\text{vertices}}^{\ell+1}$ while the variance cannot diminish faster than $\sqrt{f_\ell(N_{\text{vertices}})} \leqslant N_{\text{vertices}}^{\ell/2}$. In consequence increasing $N$ is a better way to improve the accuracy of the results than increasing the size of each sampled subgraph, since the overall computation time scales only linearly with $N$.

Empirically, we confirmed the qualitative analysis outlined above as we found that a large number $N$ of samples of size $N_{\text{vertices}} \sim 20 - 30$ vertices offered the best trade-off between accuracy and computation time. Note, in practice $\tau_{20}(N_{\text{vertices}})$ is found to be between a millisecond and a tenth of a second time.

## 3. Balance from primitive orbits

### 3.1 *Background*

A *primitive orbit* [6] on a network is a cycle which contains no backtracking steps or tail, and is not a recurrence of any other cycle, e.g. $(v_0 v_1 v_2 v_0)^2$ [42]. It is important to note that this is not the same as a simple cycle—for example if $c_1$ and $c_2$ are simple cycles sharing an edge, then $c_1 c_2$ is a primitive orbit. However, primitive orbits are identical to simple cycles up to order five [7]. The (inverse) Ihara zeta function of a graph is given by [27]

$$\zeta_{|G|}^{-1}(z) = \prod_{c \in [C]} \left(1 - z^{\ell(c)}\right), \tag{3.1}$$

where $\ell(c)$ is the length of the primitive orbit and $|G|$ denotes the unsigned version of $G$. The notation $[C]$ designates the set of equivalence classes of primitive orbits, i.e. the set of primitive orbits where all starting points on the same cycle are considered equivalent. The zeta function $\zeta_{|G|}^{-1}(z)$ can be expanded into terms representing each primitive orbit; to first order the expansion is $\zeta_{|G|}^{-1}(z) = 1 - \sum_{c \in [C]} z^{\ell(c)} + O(z^6)$. Therefore, up to order 5, the zeta function corresponds to a sum over simple cycles, and non-simple cycle contamination only occurs at order 6 and higher. Because of this, the Ihara zeta function will provide a more precise measure of balance than a walk-based one.

---

[6]Unfortunately, primitive orbits are also known as "prime cycles", which is, strictly speaking, a misnomer. Indeed, primitive orbits do not satisfy the fundamental definition of a prime element, namely $p$ is prime if and only if $p|a.b \iff p|a$ or $p|b$ for all $a, b$. Ironically, the only objects obeying this definition on a graph are the simple cycles, see [15].

[7]This is true if and only if the graph has no self-loops, i.e. length 1 cycles. In the presence of such loops, we can simply remove them by replacing $\mathbf{A}$ by $\mathbf{A} - \text{Diag}(\mathbf{A})$ in the calculations of $R_{\ell \geqslant 3}$.

To approximate the balance ratios $R_\ell$ using primitive orbits, we begin by introducing a modified version of the Ihara zeta function for signed networks

$$\zeta_G^{-1}(z) = \prod_{c \in [C]} \left(1 - s(c)z^{\ell(c)}\right),$$

where $s(c)$ is the sign of the primitive orbit $c$. We now need an efficient way of evaluating the zeta function. This can be achieved using its determinant form

$$\zeta_G^{-1}(z) = \det\left(\mathbf{I} - z\mathbf{T}\right),$$

where $\mathbf{T}$ is the Hashimoto matrix of the network, also known as edge adjacency matrix [23]. Let us recall the standard definition of $\mathbf{T}$ here. We designate by $e_i$, $1 \leqslant i \leqslant |E|$, an edge of the graph equipped with a randomly chosen orientation, and by $e_i^{-1}$ the edge $e_i$ with the opposite orientation. Then $\mathbf{T}$ is the $2|E| \times 2|E|$ matrix defined as [42]:

$$\mathbf{T}_{ij} := \begin{cases} 1, & \text{if and only if the end vertex of edge } e_i \text{ is the start vertex of edge } e_j \text{ and } e_j \neq e_i^{-1} \\ 0, & \text{otherwise.} \end{cases}$$

In other words, $\mathbf{T}$ is the adjacency matrix of the oriented line graph (OLG), that is its vertices correspond to the directed edges of the original graph. In the OLG, vertices are connected if there is an allowed two-step walk along the corresponding edges in the original graph and backtracking steps are not allowed (so $ab, bc$ is allowed if $ab$ and $bc$ are edges, but $ab, ba$ is not allowed).

To incorporate the edge signs into the matrix, we use forward sign assignment. Since all terms in the zeta function are cycles, we can uniquely place the sign of edge $ab$ into any edge in $\mathbf{T}$ which begins from $ab$, and the sign of the cycle in $\mathbf{T}$ will be the same as the sign of the original cycle.

## 3.2 *Computation*

In principle, the number of primitive orbits of any length is easily determined from traces of powers of $\mathbf{T}$. We have the following result, which we prove in Appendix B.

PROPOSITION 3.1 Let $G$ be a signed directed graph, $\mathbf{T}$ its Hashimoto adjacency matrix and $N_{\text{ob};\ell}^+$ and $N_{\text{ob};\ell}^-$ be the number of positive and negative primitive orbits of length $\ell$ on $G$, respectively. Then

$$N_{\text{ob};\ell}^+ - N_{\text{ob};\ell}^- = \frac{1}{\ell} \sum_{d|\ell} \mu(\ell/d) \operatorname{Tr} \mathbf{T}^d,$$

where $\mu(.)$ is the number-theoretic Möbius function. A similar result holds for $N_{\text{ob};\ell}^+ + N_{\text{ob};\ell}^-$ upon replacing $\mathbf{T}$ by $|\mathbf{T}|$.

This formula is particularly revealing as to the connection between the Hashimoto matrix and the primitive orbits. Traces of powers of $\mathbf{T}$ count all cycles in the OLG, i.e. the backtrackless closed walks, including the so-called power orbits, e.g. $(c_1)^2$ and $(c_1 c_2)^2$, which are not primitive. The set of cycles of length $\ell$ contains such power orbits if and only if there are orbits whose length is a divisor of $\ell$. These divisors are removed via a Möbius inversion in the above sum. For moderately sized graphs, it is straightforward to compute $\mathbf{T}$ and its spectrum, and so to compute the number of primitive orbits of any length. However, since $\mathbf{T}$ is the adjacency of the OLG, its size is equal to the number of edges in the

original network. In practice, it can therefore become difficult to compute the eigenvalues of this matrix. Since $\operatorname{Tr}\mathbf{T} = \operatorname{Tr}\mathbf{T}^2 = 0$, we need only concern ourselves with third and higher powers of the eigenvalues. The spectrum can therefore be effectively truncated, considering only the largest magnitude eigenvalues, thereby reducing the computational burden. Nevertheless, this can be computationally demanding for large networks.

For undirected unsigned graphs, Stark and Terras [41] provide a way to compute the number of all orbits (i.e. $\operatorname{Tr}\mathbf{T}^\ell$), which we adapted to count the number of positive and negative primitive orbits in a signed but *undirected* network. Let $\mathbf{D}$ be the diagonal degree matrix of $|G|$ and let $\mathbf{Q} = \mathbf{D} - \mathbf{I}$. Further let $\mathbf{A}^+$ be the adjacency of only the positive edges in $G$ and similarly $\mathbf{A}^-$ be the adjacency of only the negative edges in $G$ (coded as -1). Then the following iteration counts the number of positive and negative backtrackless closed walks (orbits) between any two vertices, denoted $W^+_{\mathrm{ob};\ell}$ and $W^-_{\mathrm{ob};\ell}$,

$$
\begin{aligned}
\mathbf{A}_2^+ &= \mathbf{A}^+\mathbf{A}^+ + \mathbf{A}^-\mathbf{A}^- - (\mathbf{Q}+\mathbf{I}), \\
\mathbf{A}_2^- &= \mathbf{A}^-\mathbf{A}^+ + \mathbf{A}^+\mathbf{A}^-, \\
\mathbf{A}_\ell^+ &= \mathbf{A}_{\ell-1}^+\mathbf{A}^+ + \mathbf{A}_{\ell-1}^-\mathbf{A}^- - \mathbf{A}_{\ell-2}^+\mathbf{Q}, \\
\mathbf{A}_\ell^- &= \mathbf{A}_{\ell-1}^-\mathbf{A}^+ + \mathbf{A}_{\ell-1}^+\mathbf{A}^- - \mathbf{A}_{\ell-2}^-\mathbf{Q}, \\
W^+_{\mathrm{ob};\ell} &= \operatorname{Tr}\left(\mathbf{A}_\ell^+ - (\mathbf{Q}-\mathbf{I})\sum_{j=1}^{(\ell-1)/2}\mathbf{A}_{\ell-2j}^+\right), \\
W^-_{\mathrm{ob};\ell} &= \operatorname{Tr}\left(\mathbf{A}_\ell^- - (\mathbf{Q}-\mathbf{I})\sum_{j=1}^{(\ell-1)/2}\mathbf{A}_{\ell-2j}^-\right).
\end{aligned}
$$

Since $W^+_{\mathrm{ob};\ell} - W^-_{\mathrm{ob};\ell} = \operatorname{Tr}\mathbf{T}^\ell$ and $W^+_{\mathrm{ob};\ell} + W^-_{\mathrm{ob};\ell} = \operatorname{Tr}|\mathbf{T}|^\ell$, using these results in conjunction with Proposition 3.1 counts the primitive orbits. This method is very efficient thanks to its use of $\mathbf{A}$ rather than $\mathbf{T}$, but remains limited to undirected networks.

Armed with the number of positive and negative primitive orbits of length $\ell$, we compute the primitive-orbits-based ratios $R^{\mathrm{ob}}_{\ell \geqslant 3} := N^-_{\mathrm{ob};\ell}/(N^-_{\mathrm{ob};\ell} + N^+_{\mathrm{ob};\ell})$. We will see that these provide better approximations of the true balance ratios $R_\ell$ than walk-based ones $R^{\mathrm{walks}}_\ell$.

## 4. Results

In this section we present the evolution of the balance ratios $R_\ell$ with $\ell$ on several social networks. As detailed in §1.4, we aim, on each network, at determining the evolution of $R_\ell$ with $\ell$. We will see that on all the large social networks studied here, $R_\ell$ exhibits a sharp transition around $\ell \sim 10 - 12$, which is absent from synthetic networks. We discuss the relation between this observation and past studies in §4.7, concluding with a possible structural interpretation for it. Tables with full numerical results for $R_\ell$, $K_\ell$ and $\mathscr{U}_\ell$ are included in Appendix C.

### 4.1 *Data sets*

Following the precedent studies by Facchetti *et al.* [14], Estrada and Benzi [13] and Aref and Wilson [3], we have analysed four social networks: i) Gahuku-Gama with 16 vertices [44]; ii) WikiElections

with 8297 vertices [46]; iii) Slashdot with 82,144 vertices [40]; and iv) Epinions with 131,828 vertices [12]. Note that among these, only the Gahuku-Gama network is undirected.

### 4.2  *Null-hypothesis*

In order to meaningfully determine if social networks are balanced, we compare our results to the balance that would be obtained on a graph with the same proportion $p$ of negative directed edges than the real network under study, but where the sign of any directed edge is negative with probability $p$. In particular, in the null-hypothesis model, the signs of any two directed edges are *independent* random variables. Then the probability that a simple cycle $c$ of length $\ell$ be negative is

$$\text{Prob}(c \text{ negative}) = \sum_{i=0}^{\lceil \ell/2 \rceil - 1} \binom{\ell}{i} p^{2i+1}(1-p)^{\ell-2i-1}. \tag{4.1}$$

Supposing for simplicity that the signs of any two simple cycles are independent random variables then the probability distribution for $N_\ell^- / (N_\ell^- + N_\ell^+)$ in the null-hypothesis is a binomial law with expectation value $R_\ell^{\text{null}}$ given by Eq. (4.1). Consequently, in this simple model the null-hypothesis is compatible up to a near 95% confidence level with any value of $R_\ell$ within the $2\sigma$ interval

$$R_\ell^{\text{null}} \pm 2 \frac{\sqrt{R_\ell^{\text{null}}(1 - R_\ell^{\text{null}})}}{\sqrt{N_\ell^- + N_\ell^+}}. \tag{4.2}$$

The assumption that the signs of any two simple cycles are independent random variables is not true on real social networks. Calculating the null-hypothesis without this assumption is very difficult in practice however. Indeed, a more accurate null model is given by evaluating the average balance ratios of all lengths over all random shufflings of the edges-signs from the social network under study. We implemented this more accurate model on the WikiElections network and found it to yield null balance ratios that are up to 9% lower than the values predicted by Eq. (4.1) when $\ell \lesssim 10$, while differences diminish for longer simple cycles. Yet, all the conclusions that can be drawn from comparing the simple null model Eqs. (4.1, 4.2) with the computed balance ratios are unchanged, since the relative positions of the two are unaltered by the more accurate model.

### 4.3  *Gahuku-Gama network*

The Gahuku-Gama network represents the relation between sixteen tribes living in the eastern central highlands of New-Guinea [17]. Since the network is very small, the Monte Carlo approach is not necessary and we obtained the exact balance from simple cycles of all lengths thanks to Eq. (2.1). The results are shown on Fig. (2). Observe how the balance ratios $R_\ell^{\text{walks}}$ and $R_\ell^{\text{ob}}$ respectively calculated from the walks and primitive orbits overestimate the proportion of balanced cycles.

The exact results show that up to length $\ell = 7$, the actual ratio $R_\ell$ is well below that of the null-hypothesis, indicating strong inter-edges correlation in favour of balanced cycles. This observation can be made more precise on noting that the balance is well fitted by a simple exponential model

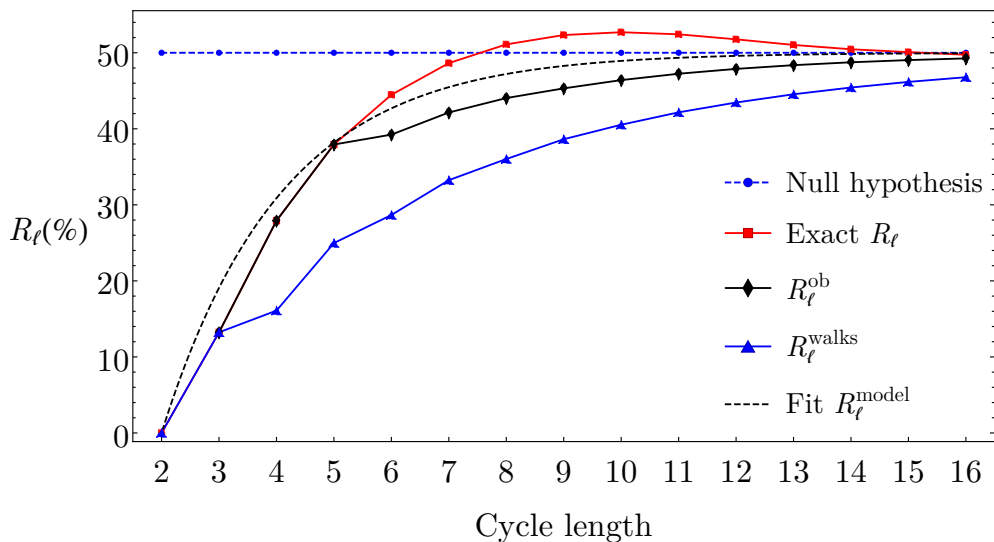$$R_\ell^{\text{model}} = \left(1 - e^{-(\ell-2)/2\xi}\right), \tag{4.3}$$

FIG. 2. Exact percentage $R_\ell$ of negatively signed simple cycles on the Gahuku-Gama network calculated from Eq. (2.1) (red squares) as compared to the null-hypothesis (blue circles). The dashed black curve is a simple exponential fit of $R_\ell$ yielding the correlation length $\xi \simeq 2.08$. Also shown are the percentages $R_{\text{walks};\ell}$ (blue triangles) and $R_{\text{ob};\ell}$ (black diamonds) calculated from the walks and primitive orbits, respectively.

where $\xi \simeq 1.04$ is the correlation length. In this context, $\xi$ characterises the distance between two edges of the graph such that their signs are correlated. Note, the maximum distance between any two vertices on a cycle of length $\ell$ is $\lfloor \ell/2 \rfloor$, hence Eq. (4.3) fits $2\xi$. This indicates that tribes of the Gahuku-Gama network are mostly sensitive to the relations with all their first degree neighbours. Furthermore, while the network is less balanced than might seem to be the case when considering only the triangles, Fig. (2) suggests that much of the imbalance is shifted to long-length simple cycles. In particular, the rebound of $R_\ell$ above 50% for $7 < \ell \lesssim 13$ shows that long negative simple cycles are over-represented as compared to a totally uncorrelated model (the null hypothesis). This means that negatively signed situations are slightly favoured in cyclic groups involving $7 < \ell \lesssim 13$ tribes. We will discuss the significance of this observation in the Section 4.7.

### 4.4   *WikiElections network*

The WikiElections network represents the votes of Wikipedia users during the elections of other users to adminship. The network is obtained as follows: when a user votes against the candidate, an edge with a negative weight is created from the voting user to the candidate. If instead the user is neutral or supports the candidate, a positive weight is given to this edge. The network counts 8,297 vertices and is thus too large for a direct exact calculation of the balance ratios beyond $\ell = 6$ and we employed a Monte Carlo approach in tandem with Eq. (2.1). In total we evaluated the balance on 1,800,000 graphs on 20 vertices. The results are shown on Fig. (3).

We find the balance ratio $R_\ell$ to evolve with $\ell$ in three major phases, which we will also observe on
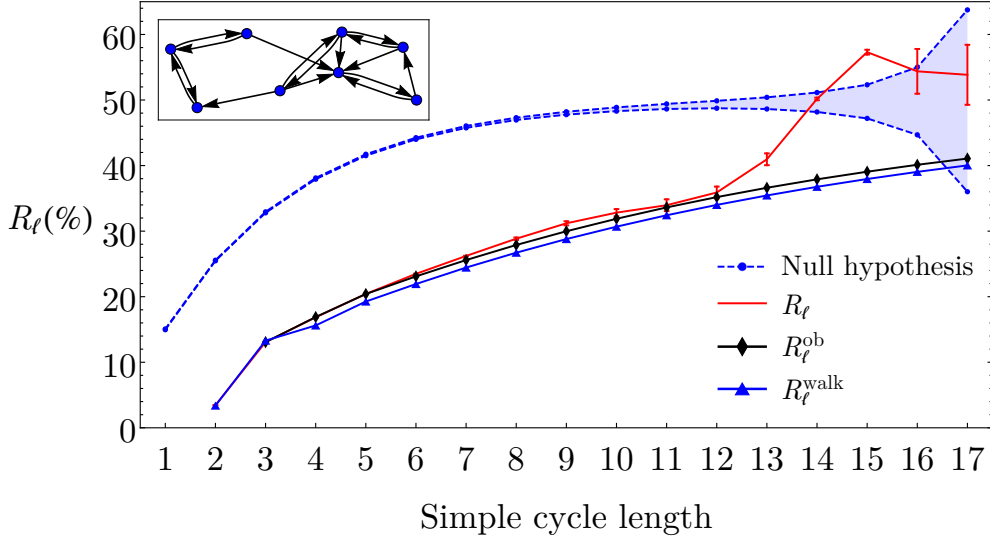
FIG. 3. Computed percentage of negatively signed simple cycles on the WikiElection network for cycle length up to 17 (red line and error bars). The blue shaded region bordered by dashed blue lines shows the values of $R_\ell$ compatible with the null-hypothesis, as determined by Eqs. (4.1, 4.2). Also shown are the percentages $R_\ell^{\mathrm{walks}}$ (blue triangles) and $R_\ell^{\mathrm{ob}}$ (black diamonds) calculated from the walks and primitive orbits, respectively. In inset, a subgraph of the WikiElections network.

the Slashdot and Epinions networks. For short simple cycles $\ell \lesssim 12$, $R_\ell$ increases slowly and smoothly with $\ell$ and is also well approximated by the primitive orbits results. In addition, within this range of cycle lengths, $R_\ell$ is always much smaller than predicted by the null-hypothesis, witnessing a strong inter-edge correlation in favour of balance. A sharp transition to $R_\ell$ values circa 50% then occurs around $\ell \sim 12 - 14$. This transition is not an artefact of our algorithm as it is not observed on synthetic networks, see Appendix A for details.

The transition demonstrates that at $\ell \sim 12 - 14$, the simple cycle length becomes longer than twice the inter-edge correlation length $\xi$, which must thus be around $6 - 7$.[8] Following the sharp transition, $R_\ell$ is reliably found to be over 50%, only to slowly decay to results consistent with null-hypothesis[9]. This last behaviour, which is also present on the Gahuku-Gama network, suggests that much of the imbalance is shifted to long simple cycles for which edges signs appear to be weakly correlated in favour of imbalance.

---

[8]Indeed, although the network is directed, many of its edges are bidirectional so that the maximum distance between any two vertices on a cycle of length $\ell$ is around $\lfloor \ell/2 \rfloor$. We emphasise that $\xi = 6 - 7$ does not mean that individuals participating in the WikiElections network are sensitive to all the relations between their neighbours up the 6th or 7th degree. Rather, $\xi$ only provides an upper bound on the *depth* of the correlation. This is because simple cycles of length $\ell$ typically sustain shortcuts which lower the average distance between the individuals participating in the cycle. The inset of Fig. 3 illustrates this phenomenon with a subgraph of the WikiElections network sustaining an octagon, but where the average distance between any two vertices is only $\sim 2.5$.

[9]This effect becomes even more pronounced when $R_\ell$ is compared with the more accurate null model that takes the structural correlations between simple cycles into account.
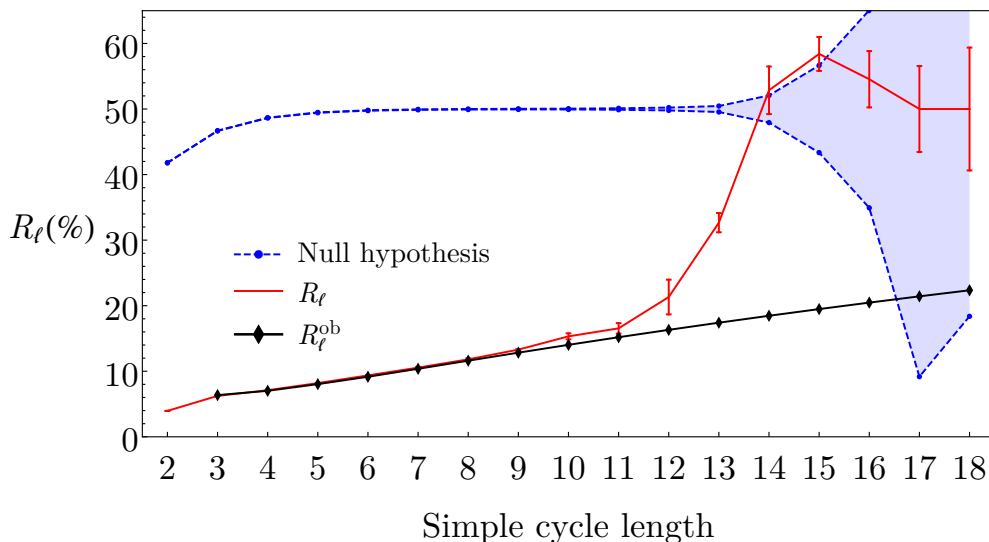
FIG. 4. Computed percentage of negatively signed simple cycles on the Slashdot network for cycle length up to 18 (red line and error bars). The blue shaded region bordered by dashed blue lines shows the values of $R_\ell$ compatible with the null-hypothesis, as determined by Eqs. (4.1, 4.2). Also shown is the percentage $R_\ell^{\mathrm{ob}}$ determined from the primitive orbits of the graph (black diamonds).

### 4.5   Slashdot network

The Slashdot network is a large directed graph on 82,144 vertices representing relations of amity/enmity between the users of the Slashdot website [31, 32]. For the Monte Carlo implementation of Eq. (2.1), we sampled 20,000,000 graphs on 20 vertices from this network. We present the balance ratio $R_\ell$ up to $\ell = 20$ on Fig. (4).

The balance ratio on this network exhibit a behaviour similar to that observed on the WikiElection network: at first $R_\ell$ increases smoothly with $\ell \leqslant 10$. Then $R_\ell$ undergoes a rapid transition to higher values broadly consistent with the null hypothesis. This indicates a correlation length between the edges of $\xi \simeq \lfloor 11/2 \rfloor = 5$ and thus a correlation depth of 5 or less. It is also remarkable that the balance ratios for $14 \leqslant \ell \leqslant 16$ are once more notably higher than 50%, indicating, as in the Gahuku-Gama and WikiElections networks, that much of the imbalance is shifted to long-length cycles.

### 4.6   Epinions

The Epinions network is a large directed graph on 131,828 vertices representing relations between the users of the consumer review website Epinions.com. For the Monte Carlo implementation of Eq. (2.1), we sampled 1,000,000,000 (one billion) graphs on 30 vertices from this network. We present the resulting balance ratio $R_\ell$ up to $\ell = 15$ on Fig. (5). We were not able to compute the balance ratio $R_\ell^{\mathrm{ob}}$ using primitive orbits, owing to the very large size of this network. Indeed, recall that in order to compute the number of primitive orbits, one needs the spectrum of the Hashimoto edge adjacency matrix of the network. This matrix has size $|E| \times |E|$, with $|E|$ the number of (directed) edges of the graph. Counting the primitive orbits thus takes $O(|E|^3)$ time and on Epinions $|E| = 841372$. Preliminary calculations indicate that it would take well over a week of computation to get sufficiently many eigenvalues to obtain a

reliable approximation of $R_\ell^{\mathrm{ob}}$. For this reason, we decided not to seek the primitive orbits results in this case.

Broadly speaking, the balance ratio $R_\ell$ behaves similarly on this network as it does on the WikiElections and Slashdot ones. The transition of $R_\ell$ from small to high values indicates a correlation length $\xi$ circa $10/2 = 5$, and thus a correlation depth of 5 or less. Strikingly, for $4 \leqslant \ell \leqslant 9$, $R_\ell$ is almost constant around 15% witnessing a very strong, almost length-independent, inter-edge correlation.

### 4.7  *Comparison with past studies and interpretation of the results*

Past studies of the WikiElections, Slashdot and Epinions networks have shown them to be very well balanced or quite unbalanced, depending on the measure studied. Kunegis [30] found that these three networks were globally well balanced, with algebraic conflicts on the order of $10^{-3}$ that is a hundred time smaller than in the case of Gama-Gahuku. Kunegis further confirmed this with the signed clustering and relative signed clustering coefficients, which are sensitive to local structures at the level of triads, and were found to be at least five times smaller on WikiElections, Slashdot and Epinions than on Gama-Gahuku. Similar conclusions had been reached by Facchetti *et al.* [14], who showed using the frustration index that the percentage of perfectly balanced nodes was well over 50% on the three online social networks.

At the opposite, Estrada and Benzi [13] who studied the degree of balance, argued that these networks are less-well balanced than concluded so far.[10] In particular they found highly unbalanced subnetworks of individuals in WikiElections, Slashdot and Epinions, which we can put in relation with the results presented here. Indeed, Estrada and Benzi predict "the existence of many negative cycles responsible for the global lack of balance" in these networks. They further remark that finding and studying these negative cycles would be extremely expensive computationally. It is tempting to relate the negative cycles predicted by Estrada and Benzi and the cyclic groups of over circa 10 individuals responsible for the overrepresentation of unbalanced simple cycles at such lengths as visible in our results.

This suggests a possible structural interpretation for the sharp transition of $R_\ell$ and the ensuing overrepresentation of unbalanced cycles that we have observed. Consider a network comprising well balanced, relatively dense clusters of individuals with negative edges mostly located in between these clusters. Then most short simple cycles will exist inside the dense clusters, consequently appearing strongly balanced. At the opposite, simple cycles whose lengths exceed the typical cluster size, being simple, are forced to cross cluster boundaries, thereby appearing mostly unbalanced. As a corollary, the highly unbalanced subnetworks of individuals are those that include individuals from different well balanced clusters and are visited mostly by long simple cycles. This kind of cluster structure can be confirmed in the case of the Gama-Gahuku network. There, Kunegis' in-depth analysis has shown that the network comprises three dense, perfectly balanced clusters with negative edges only between them. The average size of these clusters is 5.3 with the largest one regrouping 7 tribes, in line with our analysis which shows that $R_\ell \gtrsim 50\%$ from $\ell = 7$ onwards.

---

[10]The difference of conclusion might originate from that the degree of balance is sensitive to cycles of all lengths, when the signed clustering and relative signed clustering coefficients are only influenced by triads. Estrada and Benzi argued that a similar observation could be made for the frustration index, which appeared to follow closely what could be predicted from triads only.
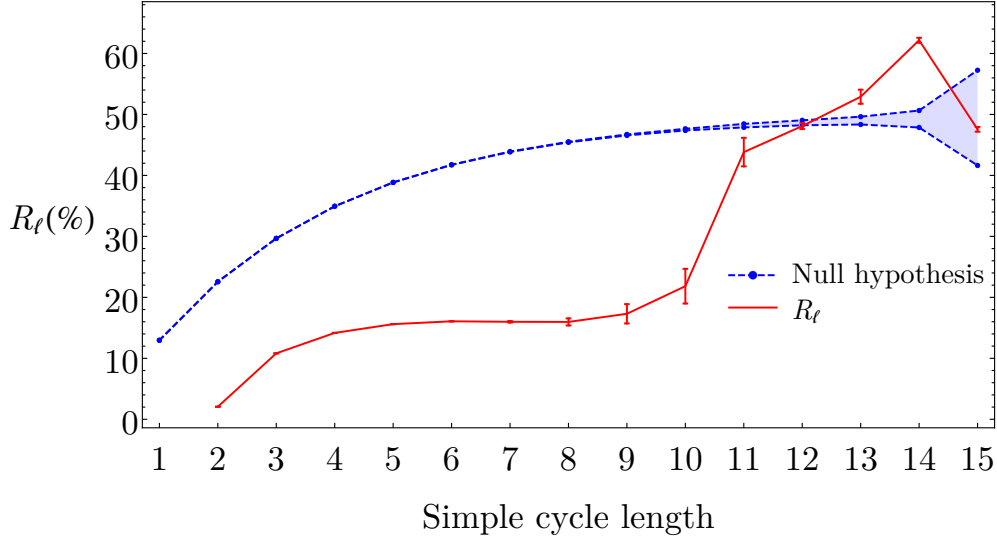
FIG. 5. Computed percentage of negatively signed simple cycles on the Epinions network for cycle length up to 15 (red line and error bars). The blue shaded region bordered by dashed blue lines shows the values of $R_\ell$ compatible with the null-hypothesis, as determined by Eqs. (4.1, 4.2).

## 5. Conclusion

### 5.1  *Balance*

Ever since relations of amity and enmity between individuals have been modelled as positive and negative signs attached to the edges of a network, the possibility of finding patterns in these signs has arisen. In one of the earliest formulation of the theory of balance, Heider proposed that most groups of individuals whose relationships form a cycle, should evolve into a state where this cycle has a positive overall sign [24]. While much research on the subject has appeared since then, altering the concept of social balance and of Heider's initial conjecture as well as questioning its validity, this very earliest statement remained largely unverified owing to its natural formulation in terms of simple cycles. Contrary to closed walks, that is "ordinary" cycles, simple cycles may not visit any vertex more than once and most problems concerning them are difficult, indeed NP-hard and #P-complete.

We thus saw Heider's earliest conjecture as a well posed challenge of network analysis and while its relevance to modern sociology is open for debate, simple cycles undeniably offer several advantages over other cycle-based studies. They are indeed free from the double-counting problem, work on undirected and directed networks alike, and the resulting balance indicators $K_\ell$ and $R_\ell$ take on a range of values, offering a gray-scale measure of the state of the network. Furthermore, distinguishing the contributions of the cycles based on length bridges the divide between local and global balance indices, since e.g. doing so probes edge-correlations on a wide range of length scales. Consequently we sought, as network analysts, to provide the most detailed possible study of social balance using hitherto unaccessible objects, the simple cycles, via novel combinatorial techniques. We hope that our results will stimulate further research into simple cycles on the one hand, and will be exploited by sociologists in the light of the latest developments in this field on the other.

We have studied the proportion $R_\ell$ of negative simple cycles of length $\ell \leqslant 20$ and its variants $K_\ell$ and $\mathcal{U}_\ell$; so as to be as close as possible to Heider's original conjecture, while also relating our results to the latest research. For exemple, $1 - R_3$ equals the triangle index and $K_3$ is the relative signed clustering coefficient. Our results show that $R_\ell$ is greatly depressed as compared to an independent sampling scenario (the null hypothesis), typically up to lengths of circa $\ell \sim 10$. This means that balanced situations are strongly favoured when short simple cycles are considered. It is interesting that on the three large networks analysed here (Wikielections, Slashdot and Epinions), a rapid transition from balance (small $R_\ell$ values) to random ($R_\ell \sim R_\ell^{\mathrm{null}}$) occurs around $\ell \sim 10$. This is a signature of strong inter-edges correlations with correlation length $\xi \simeq 10/2 = 5$. The correlation depth, which quantifies the degree up to which individuals are correlated with their neighbours, is thus less than or equal to 5. A rebound of the balance ratio over 50% following the transition is also clearly detectable in the data. This means that unbalanced situations are over-represented for long-length simple cycles as compared to the null-hypothesis.

While the relevance of cyclic clusters of individuals in social networks beyond triads is currently controversial, we believe that these do contain sociological information: we observe there is no *mathematical* reason why $R_\ell$ should undergo a sharp transition as seen in the data. In particular, this transition does not exist in the null-hypothesis model, where edge signs are attributed randomly to a real social network, nor does it exist on synthetic networks where both edges and their signs are random. We therefore hope that researchers with a more sociological background than ours will consider the results of this work.

## 5.2 *Functions on simple cycles and simple paths*

Technically, the approach presented here to study the balance in networks is generally applicable to estimate any function of the simple cycles of a graph. Furthermore, the core combinatorial result of our method immediately extend to vertex-specific questions, e.g. for evaluating the balance of the simple cycles passing through some specified vertex. It also remains valid when asking questions pertaining to simple paths (also known as self-avoiding walks). Both of these observations stem from a matrix extension of Eq. (2.1) which is presented in [16]. This extension provides a matrix $\mathbf{P}(z)$ whose entry $\mathbf{P}(z)_{ij}$ is the ordinary generating function of the simple paths from $i$ to $j$ ($i \neq j$) or of the simple-cycles from $i$ to itself ($i = j$). Note, the matrix extension is *not* obtained upon just removing the trace from Eq. (2.1).

This matrix formulation should permit the calculation of such functions as the all-paths kernel [7] for data-mining algorithms. The all-paths kernel is a positive semi-definite function on pairs of graphs aimed at measuring their degree of similarity via the similarity between their simple paths:

$$k_{\text{all paths}}(G, G') = \sum_{p \text{ simple path on } G} \sum_{p' \text{ simple path on } G'} k(p, p'),$$

with $k(p, p')$ a kernel on pairs of paths. Another example is the cyclic pattern kernel [25] which evaluates the similarities between the simple cycles of a pair of graphs. Both of these examples currently suffer from efficiency problems which one should overcome (at least partially) with Eq. (2.1) and its matrix extension, possibly in conjunction with a Monte Carlo approach as effected here.

**Acknowledgment**

## A. Checking the method against bias

While the convergence of our calculations can be assessed via the standard deviation of the results, Monte Carlo approaches are not immune to bias. It is thus necessary to verify the quality of the results independently of the method itself:

i) We computed the exact balance ratios $R_1$, $R_2$ and $R_3$ via conventional means and verified our results to be consistent with these, see Appendix C, Tables A.4, A.7 and A.10. In the case of WikiElections, we also computed $R_4$, $R_5$ and $R_6$ exactly using Eq. (2.1).

ii) When available, we used the primitive orbits results to verify that the balance ratios $R_4$ and $R_5$ predicted by the algorithm were consistent with the exact results. Indeed, recall that up to $\ell = 5$, $R_\ell^{\mathrm{ob}} = R_\ell$ exactly.

iii) On the Gahuku-Gama network, we verified that the Monte Carlo results are consistent with the exact balance ratios at all lengths.

In addition, we found the WikiElections, Slashdot and Epinions networks to exhibit sharp transitions of their balance ratios from low values $R_\ell \sim 10-30\%$ up to values consistent with the null-hypothesis circa 50%. Given the importance of this observation, it is necessary to check that it is not an artifact of the algorithm we employed:

i) In the case of the WikiElections network, we reallocated the edge signs randomly and ran our approach on the resulting signed graph. The balance ratios did not exhibit any sharp transition anymore but rather were consistent with the more accurate null model. This indicates that the transition is not an artifact of our method.

ii) We should expect a transition of $R_\ell$ to values consistent with the null hypothesis as the cycle length becomes longer than the correlation length. Indeed, if there was no sharp transition, a simple extrapolation of the trend exhibited by the first five (exactly known) balance ratios, suggests that $R_\ell$ would not be reach 50% until at least $\ell \gtrsim 50$. This conservative estimate would mean that $\xi \gtrsim 25$ or more, a number that is far too large to be plausible.

## B. Proof of Proposition 3.1

The proposition results from equating the product and determinant forms of the Ihara zeta function. Recall that

$$\zeta_{|G|}^{-1}(z) = \prod_{c \in [C]} \left(1 - z^{\ell(c)}\right) = \prod_\ell \left(1 - z^\ell\right)^{N_{\mathrm{ob};\ell}},$$

where $N_{\mathrm{ob};\ell} = N_{\mathrm{ob};\ell}^+ + N_{\mathrm{ob};\ell}^-$ is the total number of primitive orbits of length $\ell$. Thus we have

$$\zeta_{|G|}^{-1}(z) = \prod_\ell \left(1 - z^\ell\right)^{N_{\mathrm{ob};\ell}} = \det\left(\mathbf{I} - z|\mathbf{T}|\right). \tag{A.1}$$

Now, on taking the logarithm on both sides we obtain

$$\sum_{i=1}^{\infty} \frac{1}{i} z^i \operatorname{Tr} |\mathbf{T}|^i = \sum_{j}^{\infty} \sum_{k=1}^{\infty} N_{\mathrm{ob};j} \frac{z^{kj}}{k}.$$

Equating the coefficient of $z^{\ell}$ on left and right hand sides then gives

$$\frac{1}{\ell} \operatorname{Tr} |\mathbf{T}|^{\ell} = \sum_{k|\ell}^{\infty} \frac{1}{k} N_{\mathrm{ob};\ell/k},$$

and a Möbius inversion finally provides $N_{\mathrm{ob};\ell}$

$$N_{\mathrm{ob};\ell}^{+} + N_{\mathrm{ob};\ell}^{-} = \frac{1}{\ell} \sum_{k|\ell} \mu(\ell/k) \operatorname{Tr} |\mathbf{T}|^{k}.$$

The proof is entirely similar on signed networks, where $\mathbf{T}$ replaces $|\mathbf{T}|$ and $N_{\mathrm{ob};\ell}^{+} - N_{\mathrm{ob};\ell}^{-}$ is obtained instead of $N_{\mathrm{ob};\ell}^{+} + N_{\mathrm{ob};\ell}^{-}$.                                   $\square$

## C. Full numerical results

In this section we present the full numerical results obtained on the four networks mentioned earlier, which include the balance ratios, $R_{\ell} = (N_{\ell}^{-})/(N_{\ell}^{-} + N_{\ell}^{+})$, ratios of negative to positive simple cycles $\mathscr{U}_{\ell} = N_{\ell}^{-}/N_{\ell}^{+}$ and the relative signed clustering coefficient $K_{\ell} := (N_{\ell}^{+} - N_{\ell}^{-})/(N_{\ell}^{-} + N_{\ell}^{+})$ up to $\ell = 20$. We also give the exact values of $R_{\ell}$, $\mathscr{U}_{\ell}$ and $K_{\ell}$ for the self-loops, backtracks and triangles ($\ell = 1, 2, 3$). These are respectively given by

$$R_1 = \frac{\operatorname{Tr}(|\mathbf{A}| - \mathbf{A})}{2 \operatorname{Tr} |\mathbf{A}|}, \quad R_2 = \frac{\operatorname{Tr}(|\tilde{\mathbf{A}}|^2 - \tilde{\mathbf{A}}^2)}{2 \operatorname{Tr} |\tilde{\mathbf{A}}|^2}, \quad R_3 = \frac{\operatorname{Tr}(|\tilde{\mathbf{A}}|^3 - \tilde{\mathbf{A}}^3)}{2 \operatorname{Tr} |\tilde{\mathbf{A}}|^3},$$

$$\mathscr{U}_1 = \frac{\operatorname{Tr}(|\mathbf{A}| - \mathbf{A})}{\operatorname{Tr}(|\mathbf{A}| + \mathbf{A})}, \quad \mathscr{U}_2 = \frac{\operatorname{Tr}(|\tilde{\mathbf{A}}|^2 - \tilde{\mathbf{A}}^2)}{\operatorname{Tr}(|\tilde{\mathbf{A}}|^2 + \tilde{\mathbf{A}}^2)}, \quad \mathscr{U}_3 = \frac{\operatorname{Tr}(|\tilde{\mathbf{A}}|^3 - \tilde{\mathbf{A}}^3)}{\operatorname{Tr}(|\tilde{\mathbf{A}}|^3 + \tilde{\mathbf{A}}^3)},$$

$$K_1 = \frac{\operatorname{Tr}(\mathbf{A})}{\operatorname{Tr}(|\mathbf{A}|)}, \quad K_2 = \frac{\operatorname{Tr}(\tilde{\mathbf{A}}^2)}{\operatorname{Tr}(|\tilde{\mathbf{A}}|^2)}, \quad K_3 = \frac{\operatorname{Tr}(\tilde{\mathbf{A}}^3)}{\operatorname{Tr}(|\tilde{\mathbf{A}}|^3)},$$

where $\tilde{\mathbf{A}} = \mathbf{A} - \operatorname{Diag}(\mathbf{A})$.

|  | Self-loops $R_1$ | Backtracks $R_2$ | Triangles $R_3$ | Squares $R_4$ | Pentagons $R_5$ |
|---|---|---|---|---|---|
| **Exact** |  | 0% | 13.24% | 27.92% | 37.93% |

|  | $R_6$ | $R_7$ | $R_8$ | $R_9$ | $R_{10}$ |
|---|---|---|---|---|---|
| **Exact** | 44.47% | 48.64% | 51.10% | 52.33% | 52.7% |

|  | $R_{11}$ | $R_{12}$ | $R_{13}$ | $R_{14}$ | $R_{15}$ |
|---|---|---|---|---|---|
| **Exact** | 52.43% | 51.77% | 51.03% | 50.46% | 50.09% |

|  | $R_{16}$ | $R_{17}$ | $R_{18}$ | $R_{19}$ | $R_{20}$ |
|---|---|---|---|---|---|
| **Exact** | 49.74% |  |  |  |  |

Table A.1. Exact balance ratios $R_\ell$ for $1 \leqslant \ell \leqslant 20$ obtained on the Gama network using Eq. (2.1).

|  | Self-loops $\mathscr{U}_1$ | Backtracks $\mathscr{U}_2$ | Triangles $\mathscr{U}_3$ | Squares $\mathscr{U}_4$ | Pentagons $\mathscr{U}_5$ |
|---|---|---|---|---|---|
| **Exact** |  | 0% | 15.3% | 38.7% | 61.1% |

|  | $\mathscr{U}_6$ | $\mathscr{U}_7$ | $\mathscr{U}_8$ | $\mathscr{U}_9$ | $\mathscr{U}_{10}$ |
|---|---|---|---|---|---|
| **Exact** | 80.1% | 94.7% | 104.5% | 109.8% | 111.4% |

|  | $\mathscr{U}_{11}$ | $\mathscr{U}_{12}$ | $\mathscr{U}_{13}$ | $\mathscr{U}_{14}$ | $\mathscr{U}_{15}$ |
|---|---|---|---|---|---|
| **Exact** | 110.2% | 107.3% | 104.2% | 101.9% | 100.3% |

|  | $\mathscr{U}_{16}$ | $\mathscr{U}_{17}$ | $\mathscr{U}_{18}$ | $\mathscr{U}_{19}$ | $\mathscr{U}_{20}$ |
|---|---|---|---|---|---|
| **Exact** | 99% |  |  |  |  |

Table A.2. Exact ratio of negative to positive simple cycles $\mathscr{U}_\ell$ for $1 \leqslant \ell \leqslant 20$ obtained on the Gama network using Eq. (2.1).

|  | Self-loops $K_1$ | Backtracks $K_2$ | Triangles $K_3$ | Squares $K_4$ | Pentagons $K_5$ |
|---|---|---|---|---|---|
| **Exact** |  | 1 | 0.735 | 0.442 | 0.241 |

|  | $K_6$ | $K_7$ | $K_8$ | $K_9$ | $K_{10}$ |
|---|---|---|---|---|---|
| **Exact** | 0.111 | 0.027 | $-0.022$ | $-0.047$ | $-0.054$ |

|  | $K_{11}$ | $K_{12}$ | $K_{13}$ | $K_{14}$ | $K_{15}$ |
|---|---|---|---|---|---|
| **Exact** | $-0.049$ | $-0.035$ | $-0.021$ | $-0.009$ | $-0.002\%$ |

|  | $K_{16}$ | $K_{17}$ | $K_{18}$ | $K_{19}$ | $K_{20}$ |
|---|---|---|---|---|---|
| **Exact** | 0.005 |  |  |  |  |

Table A.3. Exact relative signed cluster coefficient $K_\ell$ for $1 \leqslant \ell \leqslant 20$ obtained on the Gama network using Eq. (2.1). Recall that $-1 \leqslant K_\ell \leqslant 1$ and that $K_\ell = 1$ designates a totally balanced situation, while $K_\ell = -1$ characterises a totally unbalanced one.

| | Self-loops $R_1$ | Backtracks $R_2$ | Triangles $R_3$ | Squares $R_4$ | Pentagons $R_5$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $45.488 \pm 0.048\%$ | $3.436 \pm 0.006\%$ | $13.075 \pm 0.014\%$ | $16.862 \pm 0.098\%$ | $20.421 \pm 0.052\%$ |
| **Exact** | $45.455\%$ | $3.438\%$ | $13.068\%$ | $16.875\%$ | $20.393\%$ |

| | $R_6$ | $R_7$ | $R_8$ | $R_9$ | $R_{10}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $23.43 \pm 0.05\%$ | $26.22 \pm 0.03\%$ | $28.84 \pm 0.19\%$ | $31.2 \pm 0.31\%$ | $32.82 \pm 0.54\%$ |
| **Exact** | $23.401\%$ | | | | |

| | $R_{11}$ | $R_{12}$ | $R_{13}$ | $R_{14}$ | $R_{15}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $33.95 \pm 0.9\%$ | $35.89 \pm 0.9\%$ | $40.96 \pm 0.9\%$ | $50.16 \pm 0.19\%$ | $57.3 \pm 0.36\%$ |

| | $R_{16}$ | $R_{17}$ | $R_{18}$ | $R_{19}$ | $R_{20}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $54.38 \pm 3.41\%$ | $53.85 \pm 4.57\%$ | | | |

Table A.4. Computed balance ratios $R_\ell$ for $1 \leqslant \ell \leqslant 20$ on the WikiElections network together with twice the standard deviation exhibited by the Monte Carlo results. We found no simple cycle of length 18, 19 or 20 on this network.

| | Self-loops $\mathscr{U}_1$ | Backtracks $\mathscr{U}_2$ | Triangles $\mathscr{U}_3$ | Squares $\mathscr{U}_4$ | Pentagons $\mathscr{U}_5$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $83.45 \pm 0.17\%$ | $3.56 \pm 0.006\%$ | $15.04 \pm 0.02\%$ | $20.3 \pm 0.14\%$ | $25.7 \pm 0.08\%$ |
| **Exact** | $83.33\%$ | $3.56\%$ | $15.03\%$ | | |

| | $\mathscr{U}_6$ | $\mathscr{U}_7$ | $\mathscr{U}_8$ | $\mathscr{U}_9$ | $\mathscr{U}_{10}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $30.7 \pm 0.05\%$ | $35.5 \pm 0.06\%$ | $40.5 \pm 0.37\%$ | $45.3 \pm 0.7\%$ | $48.9 \pm 1.2\%$ |

| | $\mathscr{U}_{11}$ | $\mathscr{U}_{12}$ | $\mathscr{U}_{13}$ | $\mathscr{U}_{14}$ | $\mathscr{U}_{15}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $51.4 \pm 2\%$ | $56 \pm 2.2\%$ | $69.4 \pm 2.5\%$ | $100.7 \pm 0.8\%$ | $134.2 \pm 2\%$ |

| | $\mathscr{U}_{16}$ | $\mathscr{U}_{17}$ | $\mathscr{U}_{18}$ | $\mathscr{U}_{19}$ | $\mathscr{U}_{20}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $119 \pm 15\%$ | $116 \pm 20\%$ | | | |

Table A.5. Computed ratio of negative to positive simple cycles $\mathscr{U}_\ell$ for $1 \leqslant \ell \leqslant 20$ on the WikiElections network together with twice the standard deviation exhibited by the Monte Carlo results. We found no simple cycle of length 18, 19 or 20 on this network.

| | Self-loops $K_1$ | Backtracks $K_2$ | Triangles $K_3$ | Squares $K_4$ | Pentagons $K_5$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $0.0902 \pm 0.0014$ | $0.9312 \pm 0.0001$ | $0.7385 \pm 0.0004$ | $0.663 \pm 0.003$ | $0.592 \pm 0.001$ |
| **Exact** | $0.0909$ | $0.9312$ | $0.7386$ | | |

| | $K_6$ | $K_7$ | $K_8$ | $K_9$ | $K_{10}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $0.53 \pm 0.001$ | $0.476 \pm 0.001$ | $0.423 \pm 0.005$ | $0.376 \pm 0.009$ | $0.344 \pm 0.015$ |

| | $K_{11}$ | $K_{12}$ | $K_{13}$ | $K_{14}$ | $K_{15}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $0.321 \pm 0.025$ | $0.282 \pm 0.025$ | $0.181 \pm 0.025$ | $-0.003 \pm 0.005$ | $-0.146 \pm 0.01$ |

| | $K_{16}$ | $K_{17}$ | $K_{18}$ | $K_{19}$ | $K_{20}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $-0.088 \pm 0.096$ | $-0.077 \pm 0.129$ | | | |

Table A.6. Computed relative signed cluster coefficient $K_\ell$ for $1 \leqslant \ell \leqslant 20$ on the WikiElections network together with twice the standard deviation exhibited by the Monte Carlo results.

| | Self-loops $R_1$ | Backtracks $R_2$ | Triangles $R_3$ | Squares $R_4$ | Pentagons $R_5$ |
|---|---|---|---|---|---|
| **Monte Carlo** | | $3.9995 \pm 0.001\%$ | $6.352 \pm 0.017\%$ | $7.108 \pm 0.013\%$ | $8.192 \pm 0.013\%$ |
| **Exact** | | $4.0003\%$ | $6.3608\%$ | | |

| | $R_6$ | $R_7$ | $R_8$ | $R_9$ | $R_{10}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $9.37 \pm 0.06\%$ | $10.55 \pm 0.12\%$ | $11.82 \pm 0.15\%$ | $13.3 \pm 0.09\%$ | $15.32 \pm 0.46\%$ |

| | $R_{11}$ | $R_{12}$ | $R_{13}$ | $R_{14}$ | $R_{15}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $16.5 \pm 0.8\%$ | $21.3 \pm 2.6\%$ | $32.7 \pm 1.5\%$ | $52.9 \pm 3.6\%$ | $58.4 \pm 2.6\%$ |

| | $R_{16}$ | $R_{17}$ | $R_{18}$ | $R_{19}$ | $R_{20}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $54.5 \pm 4.3\%$ | $50.0 \pm 6.6\%$ | $50.0 \pm 9.4\%$ | | |

Table A.7. Computed balance ratios $R_\ell$ for $1 \leqslant \ell \leqslant 20$ on the Slashdot network together with twice the standard deviation exhibited by the Monte Carlo results. We found no simple cycle of length 19 or 20 on this network.

| | Self-loops $\mathscr{U}_1$ | Backtracks $\mathscr{U}_2$ | Triangles $\mathscr{U}_3$ | Squares $\mathscr{U}_4$ | Pentagons $R_5$ |
|---|---|---|---|---|---|
| **Monte Carlo** | | $4.166 \pm 0.001\%$ | $6.783 \pm 0.019\%$ | $7.552 \pm 0.015\%$ | $8.923 \pm 0.015\%$ |
| **Exact** | | $4.167\%$ | $6.793\%$ | | |

| | $\mathscr{U}_6$ | $\mathscr{U}_7$ | $\mathscr{U}_8$ | $\mathscr{U}_9$ | $\mathscr{U}_{10}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $10.34 \pm 0.07\%$ | $11.79 \pm 0.15\%$ | $13.40 \pm 0.19\%$ | $15.3 \pm 0.12\%$ | $18.09 \pm 0.64\%$ |

| | $\mathscr{U}_{11}$ | $\mathscr{U}_{12}$ | $\mathscr{U}_{13}$ | $\mathscr{U}_{14}$ | $\mathscr{U}_{15}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $19.8 \pm 1.1\%$ | $27.1 \pm 4.1\%$ | $48.6 \pm 3.2\%$ | $112.3 \pm 6.7\%$ | $140 \pm 15\%$ |

| | $\mathscr{U}_{16}$ | $\mathscr{U}_{17}$ | $\mathscr{U}_{18}$ | $\mathscr{U}_{19}$ | $\mathscr{U}_{20}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $120 \pm 20\%$ | $100.0 \pm 25\%$ | $100 \pm 31\%$ | | |

Table A.8. Computed ratio of negative to positive simple cycles $\mathscr{U}_\ell$ for $1 \leqslant \ell \leqslant 20$ on the Slashdot network together with twice the standard deviation exhibited by the Monte Carlo results. We found no simple cycle of length 19 or 20 on this network.

| | Self-loops $K_1$ | Backtracks $K_2$ | Triangles $K_3$ | Squares $K_4$ | Pentagons $R_5$ |
|---|---|---|---|---|---|
| **Monte Carlo** | | $0.92001 \pm 0.00003$ | $0.8730 \pm 0.0005$ | $0.8578 \pm 0.0004$ | $0.8362 \pm 0.0004$ |
| **Exact** | | $0.92000$ | $0.8728$ | | |

| | $K_6$ | $K_7$ | $K_8$ | $K_9$ | $K_{10}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $0.813 \pm 0.002$ | $0.789 \pm 0.003$ | $0.764 \pm 0.004$ | $0.734 \pm 0.003$ | $0.694 \pm 0.013$ |

| | $K_{11}$ | $K_{12}$ | $K_{13}$ | $K_{14}$ | $K_{15}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $0.67 \pm 0.023$ | $0.574 \pm 0.074$ | $0.346 \pm 0.042$ | $-0.058 \pm 0.102$ | $-0.168 \pm 0.074$ |

| | $K_{16}$ | $K_{17}$ | $K_{18}$ | $K_{19}$ | $K_{20}$ |
|---|---|---|---|---|---|
| **Monte Carlo** | $-0.09 \pm 0.122$ | $0 \pm 0.187$ | $0.001 \pm 0.266$ | | |

Table A.9. Computed relative signed cluster coefficient $K_\ell$ for $1 \leqslant \ell \leqslant 20$ on the Slashdot network together with twice the standard deviation exhibited by the Monte Carlo results.

|  | Self-loops $R_1$ | Backtracks $R_2$ | Triangles $R_3$ | Squares $R_4$ | Pentagons $R_5$ |
|---|---|---|---|---|---|
| Monte Carlo | $6.1082 \pm 0.0002\%$ | $2.0857 \pm 0.0003\%$ | $11.2355 \pm 0.0016\%$ | $14.17 \pm 0.002\%$ | $15.61 \pm 0.007\%$ |
| Exact | $6.1082\%$ | $2.0858\%$ | $11.2343\%$ |  |  |

|  | $R_6$ | $R_7$ | $R_8$ | $R_9$ | $R_{10}$ |
|---|---|---|---|---|---|
| Monte Carlo | $16.07 \pm 0.02\%$ | $15.99 \pm 0.12\%$ | $15.97 \pm 0.58\%$ | $17.30 \pm 1.58\%$ | $21.83 \pm 2.84\%$ |

|  | $R_{11}$ | $R_{12}$ | $R_{13}$ | $R_{14}$ | $R_{15}$ |
|---|---|---|---|---|---|
| Monte Carlo | $43.8 \pm 2.3\%$ | $48.1 \pm 0.5\%$ | $52.9 \pm 1.2\%$ | $62.2 \pm 0.4\%$ | $47.6 \pm 0.4\%$ |

|  | $R_{16}$ | $R_{17}$ | $R_{18}$ | $R_{19}$ | $R_{20}$ |
|---|---|---|---|---|---|
| Monte Carlo |  | $50\%$ |  | $44.4\%$ |  |

Table A.10. Computed balance ratios $R_\ell$ for $1 \leqslant \ell \leqslant 20$ on the Epinions network together with twice the standard deviation exhibited by the Monte Carlo results. We found no simple cycle of length 16, 18 or 20 on this network. Furthermore, we were unable to determine the standard deviation of the balance for $R_{17}$ and $R_{19}$. Regardless of the accuracy on these results, the small numbers of simple cycles of such lengths that we found imply that the null-hypothesis is compatible with all values $15\% \lesssim R_\ell \lesssim 85\%$, as per Eq. (4.2).

|  | Self-loops $\mathscr{U}_1$ | Backtracks $\mathscr{U}_2$ | Triangles $\mathscr{U}_3$ | Squares $\mathscr{U}_4$ | Pentagons $\mathscr{U}_5$ |
|---|---|---|---|---|---|
| Monte Carlo | $6.5056 \pm 0.0003\%$ | $2.1301 \pm 0.0004\%$ | $12.658 \pm 0.0028\%$ | $16.51 \pm 0.004\%$ | $18.50 \pm 0.014\%$ |
| Exact | $6.5056\%$ | $2.1302\%$ | $12.656\%$ |  |  |

|  | $\mathscr{U}_6$ | $\mathscr{U}_7$ | $\mathscr{U}_8$ | $\mathscr{U}_9$ | $\mathscr{U}_{10}$ |
|---|---|---|---|---|---|
| Monte Carlo | $19.15 \pm 0.04\%$ | $19.03 \pm 0.24\%$ | $19.01 \pm 1.16\%$ | $20.96 \pm 3.27\%$ | $28.09 \pm 5.58\%$ |

|  | $\mathscr{U}_{11}$ | $\mathscr{U}_{12}$ | $\mathscr{U}_{13}$ | $\mathscr{U}_{14}$ | $\mathscr{U}_{15}$ |
|---|---|---|---|---|---|
| Monte Carlo | $78.2 \pm 10.3\%$ | $92.7 \pm 2.6\%$ | $112.5 \pm 7.7\%$ | $164.6 \pm 4.0\%$ | $90.9 \pm 2.1\%$ |

|  | $\mathscr{U}_{16}$ | $\mathscr{U}_{17}$ | $\mathscr{U}_{18}$ | $\mathscr{U}_{19}$ | $\mathscr{U}_{20}$ |
|---|---|---|---|---|---|
| Monte Carlo |  | $100\%$ |  | $79.9\%$ |  |

Table A.11. Computed ratio of negative to positive simple cycles $\mathscr{U}_\ell$ for $1 \leqslant \ell \leqslant 20$ on the Epinions network together with twice the standard deviation exhibited by the Monte Carlo results.

|  | Self-loops $K_1$ | Backtracks $K_2$ | Triangles $K_3$ | Squares $K_4$ | Pentagons $K_5$ |
|---|---|---|---|---|---|
| Monte Carlo | $0.87784 \pm 0.00001$ | $0.95829 \pm 0.00001$ | $0.77529 \pm 0.00005$ | $0.7166 \pm 0.0001$ | $0.6878 \pm 0.0002$ |
| Exact | $0.87784$ | $0.95828$ | $0.77531$ |  |  |

|  | $K_6$ | $K_7$ | $K_8$ | $K_9$ | $K_{10}$ |
|---|---|---|---|---|---|
| Monte Carlo | $0.679 \pm 0.001$ | $0.68 \pm 0.003$ | $0.681 \pm 0.016$ | $0.654 \pm 0.045$ | $0.563 \pm 0.08$ |

|  | $K_{11}$ | $K_{12}$ | $K_{13}$ | $K_{14}$ | $K_{15}$ |
|---|---|---|---|---|---|
| Monte Carlo | $0.124 \pm 0.065$ | $0.038 \pm 0.014$ | $-0.058 \pm 0.034$ | $-0.244 \pm 0.011$ | $0.048 \pm 0.011$ |

|  | $K_{16}$ | $K_{17}$ | $K_{18}$ | $K_{19}$ | $K_{20}$ |
|---|---|---|---|---|---|
| Monte Carlo |  | $0$ |  | $0.112$ |  |

Table A.12. Computed relative signed cluster coefficient $K_\ell$ for $1 \leqslant \ell \leqslant 20$ on the Epinions network together with twice the standard deviation exhibited by the Monte Carlo results.

## References

1. Alon, N., Yuster, R. & Zwick, U. (1997) Finding and counting given length cycles. *Algorithmica*, **17**, 209–223.
2. Antal, T., Krapivsky, P. L. & Red, S. (2006) Social balance on networks: The dynamics of friendship and enmity. *Physica D: Nonlinear Phenomena*, **224**, 130–136.
3. Aref, S. & Wilson, M. C. (2015) Measuring Partial Balance in Signed Networks. *arXiv:1509.04037*.
4. Avis, D. & Fukuda, K. (1996) Reverse search for enumeration. *Discrete Applied Mathematics*, **65**, 21–46.
5. Bax, E. & Franklin, J. (1996) A finite-difference sieve to count paths and cycles by length. *Information Processing Letters*, **60**(4), 171–176.
6. Birmelé, E., Ferreira, R., Grossi, R., Marino, A., Pisanti, N., Rizzi, R. & Sacomoto, G. (2013) Optimal listing of cycles and st-paths in undirected paths. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1884–1896.
7. Borgwardt, K. M. & Kriegel, H.-P. (2005) Shortest-path kernels on graphs. In *Proceedings of the 5th IEEE International Con- ference on Data Mining (ICDM 2005), 27-30 November 2005*.
8. Bürgisser, P., Clausen, M. & Shokrollahi, M. A. (1997) *Algebraic Complexity Theory*, volume 315 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin; New York.
9. Cartwright, D. & Harary, F. (1956) Structural balance: a generalization of Heider's theory. *Psychological Review*, **63**, 277–293.
10. Chiang, K.-Y., Hsieh, C.-J., Natarajan, N. & Dhillon, I. S. (2014) Prediction and Clustering in Signed Networks: A Local to Global Perspective. *Journal of Machine Learning Research*, **15**, 1177–1213.
11. Elbassioni, K. (2015) A Polynomial Delay Algorithm for Generating Connected Induced Subgraphs of a Given Cardinality. *Journal of Graph Algorithms and Applications*, **19**, 273–280.
12. Epinions social network (2016) Epinions social network. `http://snap.stanford.edu/data/soc-sign-epinions.html`.
13. Estrada, E. & Benzi, M. (2014) Walk-based measure of balance in signed networks: Detecting lack of balance in social networks. *Physical Review E*, **90**, 042802.
14. Facchetti, G., Iacono, G. & Altafini, C. (2011) Computing global structural balance in large-scale signed social networks. *Proceedings of the National Academy of Sciences*, **108**, 20953–20958.
15. Giscard, P.-L. & Rochet, P. (2016a) Algebraic Combinatorics on Trace Monoids: Extending Number Theory to Walks on Graphs. *arXiv:1601.01780*.
16. Giscard, P.-L. & Rochet, P. (2016b) Enumerating simple paths from connected induced subgraphs. *arXiv:1606.00289*.
17. Hage, P. (1973) A graph theoretic approach to the analysis of alliance structure and local grouping in highland New Guinea. *Anthropological Forum: A Journal of Social Anthropology and Comparative Sociology*, **3**, 280–294.
18. Harary, F. (1953) On the notion of balance of a signed graph. *Michigan Math. J.*, **2**, 143–146.
19. Harary, F. (1959) On the measurement of structural balance. *Behavioral Science*, **4**, 316–323.
20. Harary, F. (1960) A matrix criterion for structural balance. *Naval Research Logistics*, **7**, 195–199.
21. Harary, F. & Kabell, J. A. (1980) A simple algorithm to detect balance in signed graphs. *Mathematical Social Sciences*, **1**, 131–136.
22. Harary, F. & Manvel, B. (1971) On the number of cycles in a graph. *Matematický časopis*, **21**, 55–63.
23. Hashimoto, K. (1989) Zeta functions of finite graphs and representations of p-adic groups. *Stud. Pure Math.*, **15**, 211–280.
24. Heider, F. (1946) Attitudes and cognitive organization. *The Journal of Psychology*, **21**, 107–112.
25. Horváth, T., Gartner, T. & Wrobel, S. (2004) Cyclic pattern kernels for predictive graph mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
26. Iacono, G., Ramezani, F., Soranzo, N. & Altafini, C. (2010) Determining the distance to monotonicity of a biological network: a graph-theoretical approach. *IET Systems Biology*, **4**, 223–235.
27. Ihara, Y. (1966) On discrete subgroups of the two by two projective linear group over p-adic fields. *Journal of the Mathematical Society of Japan*, **18**, 219–235.

**28.** Impagliazzo, R. & Paturi, R. (2001) On the Complexity of *k*-SAT. *Journal of Computer and System Sciences*, **62**, 367–375.

**29.** Johnson, D. B. (1975) Finding all the elementary circuits of a directed graph. *SIAM J. Comput.*, **4**, 77–84.

**30.** Kunegis, J. (2014) Applications of Structural Balance in Signed Social Networks. *arXiv:1402.6865*.

**31.** Kunegis, J., Schmidt, S., Lommatzsch, A. S., Lerner, J., DeLuca, E. W. & Albayr, S. (2010) Spectral analysis of signed graphs for clustering, prediction and visualization. *SIAM Conference on Data Mining (Society for Industrial and Applied Mathematics, Philadelphia)*, pages 559–570.

**32.** Lampe, C. A., Johnston, E. & Resnick, P. (2007) Follow the reader: Filtering comments on Slashdot. *Proceedings of Computer/Human Interaction 2007 Conference (Association for Computing Machinery, New York)*, pages 1253–1262.

**33.** Lange, C., Liu, S., Peyerimhoff, N. & Post, O. (2015) Frustration index and Cheeger inequalities for discrete and continuous magnetic Laplacians. *Calculus of variations and partial differential equations*, **54**, 4165–4196.

**34.** Leskovec, J., Huttenlocher, D. & Kleinberg, J. (2010) Signed networks in social media. *Conference on Human Factors in Computing Systems*.

**35.** Mateti, P. & Deo, N. (1976) On Algorithms for Enumerating all Circuits of a Graph. *SIAM J. Comput.*, **5**, 90–99.

**36.** Movarraei, N. & Boxwala, S. A. (2016) On the Number of Cycles in a Graph. *Open Journal of Discrete Mathematics*, pages 41–69.

**37.** Norman, R. Z. (1972) A derivation of a measure of relative balance for social structures and a characterization of extensive ratio systems. *Journal of Mathematical Psychology*, **9**, 66–91.

**38.** P.-L. Giscard, N. Kriege, R. C. W. (2016) A general purpose algorithm for counting simple cycles and simple paths of any length. *arXiv:1612.05531*.

**39.** S. Aref, A. J. Mason, M. C. W. (2016) An exact method for computing the frustration index in signed networks using binary programming. *arXiv:1611.09030*.

**40.** Slashdot social network, February 2009 (2016) Slashdot social network, February 2009. `http://snap.stanford.edu/data/soc-sign-Slashdot090221.html`.

**41.** Stark, H. M. & Terras, A. A. (1996) Zeta functions of finite graphs and coverings. *Advances in Mathematics*, **121**, 124–165.

**42.** Terras, A. A. (2011) *Zeta Functions of Graphs: A Stroll through the Garden*. Cambridge University Press, Cambridge, 1st edition.

**43.** Terzi, E. & Winkler, M. (2011) A spectral algorithm for computing social balance. *Proceedings of the 8th international conference on Algorithms and models for the web graph*, pages 1–13.

**44.** UCINET IV Datasets (2016) UCINET IV Datasets, Read Highland Tribes. `http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm`.

**45.** Uehara, R. (1999) The number of connected components in graphs and its applications. *IEICE Technical Report, COMP99-10*.

**46.** Wikipedia adminship election data (2016) Wikipedia adminship election data. `http://snap.stanford.edu/data/wiki-Elec.html`.

**47.** Zajonc, R. B. & Burnstein, E. (1965) Structural balance, reciprocity, and positivity as sources of cognitive bias. *Journal of Personality*, **33**, 570–583.