Evaluating balance on social networks from their simple cycles

P.-L. Giscard[1], P. Rochet[2] and R. C. Wilson[1]
[1] *University of York, Department of Computer Science, UK*
[2] *Université de Nantes, Laboratoire de Mathématiques Jean Leray, France*
(Dated: June 10, 2016)

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 most natural formulation, the balance of a social network is 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 always been studied via other, less-direct means. In this article we present the balance as measured from the simple cycles and primitive orbits of social networks. We use a Monte Carlo implementation of a novel exact formula for counting the simple cycles on any weighted directed graph. We show that social networks exhibit strong inter-edge correlations favoring balanced situations and we determine the corresponding correlation length ξ. For longer simple cycles, the percentage of unbalanced simple cycles undergoes a rapid transition to values expected from an uncorrelated model. Our method is more generally applicable to evaluate arbitrary functions over the simple cycles and simple paths of any weighted directed graph and can also answer vertex-specific questions.

PACS numbers: 89.75.Fb, 89.65.Ef

I. INTRODUCTION

A. 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 there are instead enemies [1–5]. 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 [1, 2, 6]. The question of wether this holds for real-social networks and if not, by how much this fails to be true, arose from these considerations in the 1940s [1].

Mathematically speaking, this sociological question translates into the following problem: on a signed network G, determine for all ℓ the percentage of negative simple cycles of length ℓ. 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 since, for example, the problem of determining if a Hamiltonian cycle exists in a graph is NP-complete.

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 II. 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 III. The results produced by both approaches on four social networks are discussed and compared in Section IV.

B. Notation

Throughout this article, we consider signed directed networks $G = (\mathcal{V}; \mathcal{E})$, 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, $v_0v_1v_2v_0$ and $v_1v_2v_0v_1$ represent the same triangle, which is however distinct from $v_0v_2v_1v_0$. The number of positive and negative simple cycles of length ℓ on G are designated by N_ℓ^+ and N_ℓ^-,

respectively.

When discussing the balance of a network, we refer to the ratio R_ℓ of the number of negatively signed simple cycles of length ℓ to the total number of simple cycles of length ℓ, i.e.

$$R_\ell := \frac{N_\ell^-}{N_\ell^- + N_\ell^+}.$$

In particular, $R_\ell = 0$ when the network is perfectly balanced for length ℓ, while $R_\ell = 1$ indicates a totally unbalanced situation.

C. Existing approach using walks

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 [9] proposed the use of

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

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 $K = D/|D|$, the ratio of negative to positive cycles can be obtained as

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

This is an extremely efficient method which simply requires the evaluation of the eigenvalues of the adjacency matrix.

Expression (1) counts all closed walks (weighted by a factor $1/\ell!$ for a walk of length ℓ). Because of this, backtracking steps and multiple cycles are counted. For example, the non-simple cycle $v_0v_1v_2v_0v_1v_2v_0 = (v_0v_1v_2v_0)^2$ 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) contains non-simple cycles at order 2 and higher effectively mixing the balance ratios R_ℓ at all lengths. In addition, the global signature of balance $\mathcal{U}_{\text{walks}}$ 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 $\mathcal{U}^{\text{walks}} = 0.106$, 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 $D_\ell := \operatorname{Tr}\mathbf{A}^\ell$, $K_\ell := D_\ell/|D_\ell|$, $\mathcal{U}_\ell^{\text{walks}} := (1-K_\ell)/(1+K_\ell)$ and

$$R_\ell^{\text{walks}} := \frac{\operatorname{Tr}|\mathbf{A}|^\ell - \operatorname{Tr}\mathbf{A}^\ell}{2\operatorname{Tr}|\mathbf{A}|^\ell}. \tag{2}$$

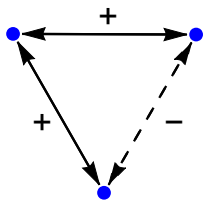

FIG. 1: An unbalanced network of three vertices. The dotted line represents a negative relationship. The exponential unbalanced ratio is $\mathcal{U}_{\text{walks}} = 0.106$ while in fact $R_2 = 1/3$ and $R_3 = 1$.

These quantities only take walks of length ℓ into account when calculating the balance. Yet, since short cycles and their multiples are typically much more abundant than long cycles, the values of $\mathcal{U}_\ell^{\text{walks}}$ and R_ℓ^{walks} are still largely dominated by the contributions from self-loops, backtracks and triangles. Consequently R_ℓ^{walks} will be depressed as compared to the true balance ratio R_ℓ, that is, R_ℓ^{walks} overestimates the proportion of balanced cycles. We demonstrate this concretely in Section IV, Figures (2) and (3), where we compare R_ℓ^{walks} with R_ℓ 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 [9, 10], 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.

II. BALANCE FROM SIMPLE CYCLES

A. Core combinatorial engine

One possible strategy to estimate the balance ratios R_ℓ 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.

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

$$P(z) = \int \frac{1}{z} \sum_{\substack{H \prec G \\ H \text{ connected}}} \mathrm{Tr}\left(\left(z\mathbf{A}_H\right)^{|H|} \left(\mathbf{I} - z\mathbf{A}_H\right)^{|N(H)|} \right) dz. \tag{3}$$

In this expression H is a *connected* induced subgraph of G [23], $\mathbf{A}_H$ 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. (3), as well as further exact formulas for $P(z)$, is presented in [7] and shall not be proven here.

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. (3) is that a brute force algorithm for finding the connected induced subgraphs of a graph, for example by breadth-first search, necessitates $O(n_H(G))$ operations. In comparison a direct search of the simple cycles themselves will require $O(n_c(G))$ operations. Eq. (3) thus represents a substantial speed-up, as can be seen on the complete graph K_N on N vertices—which is the worst case scenario. On K_N we have $n_c(G) \simeq e \times (N-1)!$, while $n_H(G) = 2^N$ is "only" exponential [24]. Furthermore, in practice, the computational cost is much smaller.

Indeed, most importantly for applications, Eq. (3) is well suited to truncations: only those connected induced subgraphs H of G for which $|H| \leq \ell \leq |H| + |N(H)|$ can possibly contribute to the coefficient of z^ℓ in $P(z)$. This means that if one is interested in the first ℓ terms of $P(z)$—that is in the simple cycles of length up to ℓ—it suffices to consider those connected induced subgraphs of G with $|H| \leq \ell$. Since furthermore only the small adjacency matrices $\mathbf{A}_H$ enter Eq. (3), each term of the equation costs $O(|H|^3) \leq O(\ell^3)$ to evaluate. Thus, when $N \gg \ell$, getting the first ℓ terms of $P(z)$ from Eq. (3) costs $O\big(\ell^3 n_H(\ell)\big) = O(\ell^3 N^\ell)$ operations in the worst case scenario, and *far less* on sparse graphs. A good rule of thumb to evaluate the computational cost on sparse graphs is as follows: let Δ be the average vertex-degree on the network. Then we can expect $N\Delta^{\ell-1}$ connected induced subgraphs on at most ℓ vertices. Thus Eq. (3) should produce $P(z)$ exactly up to order ℓ in $O\big(\ell^3 N \Delta^{\ell-1}\big)$ operations [25].

To give concrete examples, with an Intel Core i7-4790 CPU @ 3.60 GHz desktop computer, evaluating Eq. (3) on the complete graph on 15 vertices took on average $\sim 18\,\mathrm{sec}$, yielding $255,323,504,932 \simeq 2.5 \times 10^{11}$ for the total number of simple cycles, which we verify analytically to be exact. This went up to $\sim 10\,\mathrm{min}$ for the complete graph on 20 vertices where a mind-boggling $349,096,664,728,623,336 \simeq 3.5 \times 10^{17}$ simple cycles were found, 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. (3). 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.

B. 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. (3) becomes intractable and the core combinatorial engine must be supplemented by a Monte Carlo approach.

The reader may have noticed upon close inspection of Eq. (3) that $P(z)$ results from subtle cancellations between the contributions of the various connected induced subgraphs H of G. For this reason, Eq. (3) 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. (3) 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_ℓ *exactly* up to the desired length ℓ on each of these samples via Eq. (3). The average value of all the R_ℓ 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_ℓ. 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.

III. BALANCE FROM PRIMITIVE ORBITS

A. Background

A *primitive orbit* [26] on a network is a cycle which contains no backtracking steps or tail, and is not a recurrence of any other cycle, e.g. $(v_0v_1v_2v_0)^2$ [11]. 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_1c_2 is a primitive orbit. However, primitive orbits are identical to simple cycles up to order 5 [27]. The (inverse) Ihara zeta function of a graph is given by [12]

$$\zeta_{|G|}^{-1}(z) = \prod_{c\in[C]} \left(1 - z^{\ell(c)}\right), \tag{4}$$

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.

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

$$\zeta_{G}^{-1}(z) = \prod_{c\in[C]} \left(1 - s(c)z^{\ell(c)}\right), \tag{5}$$

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. This matrix is the adjacency of the oriented line graph (OLG), and so $\mathbf{T}$ has vertices corresponding to (directional) edges of the original graph. Vertices are connected if there is an allowed two-step walk along the corresponding edges in the original graph. 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.

B. 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 III.1. *Let G be a signed directed graph, $\mathbf{T}$ its Hashimoto adjacency matrix and $N^+_{ob;\,\ell}$ and $N^-_{ob;\,\ell}$ be the number of positive and negative primitive orbits of length ℓ on G, respectively. Then*

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

where $\mu(.)$ is the number-theoretic Möbius function. A similar result holds for $N^+_{ob;\,\ell} + N^-_{ob;\,\ell}$ 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 $(c_1c_2)^2$, which are not primitive. The set of cycles of length ℓ contains such power orbits if and only if there are orbits whose length is a divisor of ℓ. 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 $\mathrm{Tr}\,\mathbf{T} = \mathrm{Tr}\,\mathbf{T}^2 = 0$, 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 [13] provide a way to compute the number of all orbits (i.e. $\mathrm{Tr}\,\mathbf{T}^\ell$), 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 $W^-_{\mathrm{ob};\,\ell}$,

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

Since $W^+_{\mathrm{ob};\,\ell} - W^-_{\mathrm{ob};\,\ell} = \mathrm{Tr}\,\mathbf{T}^\ell$ and $W^+_{\mathrm{ob};\,\ell} + W^-_{\mathrm{ob};\,\ell} = \mathrm{Tr}\,|\mathbf{T}|^\ell$, using these results in conjunction with Proposition III.1 counts the primitive orbits. This method is very efficient

thanks to its use of **A** rather than **T**, but remains limited to undirected networks.

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

IV. RESULTS

A. Data sets

Following the precedent studies by Facchetti *et al.* [14] and Estrada and Benzi [9], we have analysed four social networks: i) Gahuku-Gama with 16 vertices [15]; ii) WikiElections with 8297 vertices [16]; iii) Slashdot with 82,144 vertices [17]; and iv) Epinions with 131,828 vertices [18]. Note that among these, only the Gahuku-Gama network is undirected.

B. 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 ℓ be negative is

$$\text{Prob}(c \text{ negative}) = \sum_{i=0}^{\lceil \ell/2\rceil -1} \binom{\ell}{i} p^{2i+1}(1-p)^{\ell-2i-1}. \quad (6)$$

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 $R^{\rm null}_\ell$ given by Eq. (6). Consequently, in this simple model the null-hypothesis is compatible up to a near 95% confidence level with any value of R_ℓ within the 2σ interval

$$R^{\rm null}_\ell \pm 2\frac{\sqrt{R^{\rm null}_\ell(1-R^{\rm null}_\ell)}}{\sqrt{N^-_\ell+N^+_\ell}}. \quad (7)$$

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. (6) 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. (6, 7) with the computed balance ratios are unchanged, since the relative positions of the two are unaltered by the more accurate model.

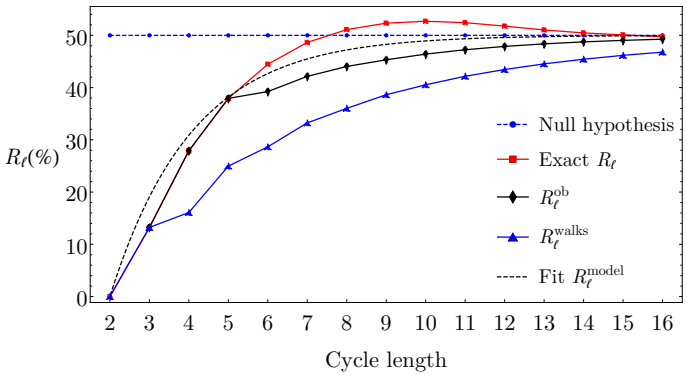

FIG. 2: Exact percentage R_ℓ of negatively signed simple cycles on the Gahuku-Gama network calculated from Eq. (3) (red squares) as compared to the null-hypothesis (blue circles). The dashed black curve is a simple exponential fit of R_ℓ yielding the correlation length $\xi \simeq 2.08$. Also shown are the percentages $R_{{\rm walks};\ell}$ (blue triangles on dashed curve) and $R_{{\rm ob};\ell}$ (black diamonds) calculated from the walks and primitive orbits, respectively.

C. Gahuku-Gama network

The Gahuku-Gama network represents the relation between sixteen tribes living in the eastern central highlands of New-Guinea [19]. Since the network is very small, the Monte Carlo approach is not necessary and we obtained the exact balance from simple cycles of all length thanks to Eq. (3). The results are shown on Fig. (2). They demonstrate how the balance ratio $R^{\rm walks}_\ell$ obtained from the walks strongly overestimates the proportion of balanced cycles. The ratio $R^{\rm ob}_\ell$ obtained from the primitive orbits suffers from the same issue, but to much lesser extend.

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

$$R^{\rm model}_\ell = \left(1 - e^{-(\ell-2)/2\xi}\right), \quad (8)$$

where $\xi \simeq 1.04$ is the correlation length. Note, the maximum distance between any two vertices on a cycle of length ℓ is $\lfloor \ell/2 \rfloor$, hence Eq. (8) fits 2ξ. This indicates that tribes of the Gahuku-Gama network are mostly sen-

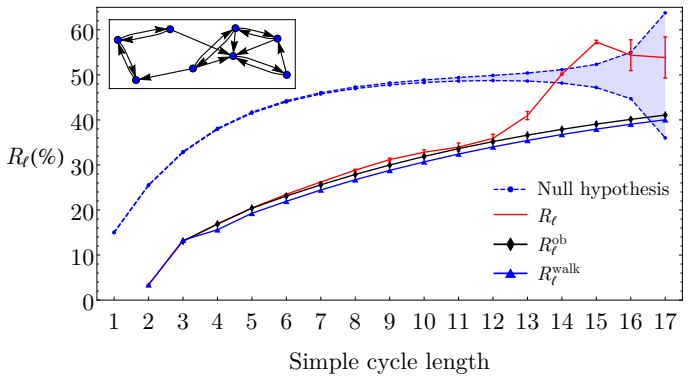

FIG. 3: Computed percentage of negatively signed simple cycles on the WikiElection network for cycle length up to 17 (red line and error bars). The blue shaded region bordered by dashed blue lines shows the values of R_ℓ compatible with the null-hypothesis, as determined by Eqs. (6, 7). Also shown are the percentages $R_\ell^{\rm walks}$ (blue triangles) and $R_\ell^{\rm ob}$ (black diamonds) calculated from the walks and primitive orbits, respectively. In inset, a subgraph of the WikiElections network.

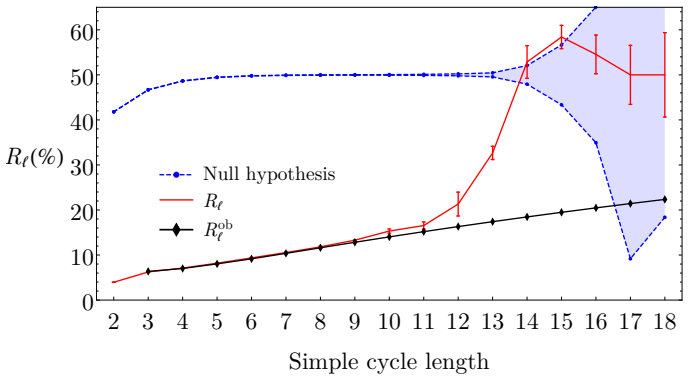

FIG. 4: Computed percentage of negatively signed simple cycles on the Slashdot network for cycle length up to 18 (red line and error bars). The blue shaded region bordered by dashed blue lines shows the values of R_ℓ compatible with the null-hypothesis, as determined by Eqs. (6, 7). Also shown is the percentage $R_\ell^{\rm ob}$ determined from the primitive orbits of the graph (black diamonds).

sitive 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) shows that much of the imbalance is shifted to long-length simple cycles. In particular, the rebound of R_ℓ above 50% for $7 < \ell \lesssim 13$ suggests that social tensions are less potent when distributed over many actors.

D. 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 ratio R_ℓ and we employed a Monte Carlo approach in tandem with Eq. (3). In total we evaluated the balance on 1,800,000 graphs on 20 vertices. The results are shown on Fig. (3) and given in full in Appendix C, Table II.

We find the balance ratio R_ℓ to evolve with ℓ in three major phases. Remarkably, we will see that these phases are also present on the Slashdot and Epinions networks. For short simple cycles $\ell \lesssim 12$, R_ℓ increases slowly and smoothly with ℓ and is also well approximated by the primitive orbits results. In addition, within this range of cycle lengths, R_ℓ is always much smaller than predicted by the null-hypothesis, witnessing a strong inter-edge correlation in favor of balance. A sharp transition to R_ℓ values circa 50% then occurs around $\ell \sim 12 - 14$. We verified that this transition is not an artifact of our algorithm, 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 ξ, which must thus be around $6 - 7$. 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 ℓ 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, ξ only provides an upper bound on the *depth* of the correlation. This is because simple cycles of length ℓ 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 ~ 2.5.

Following the sharp transition, R_ℓ is reliably found to be over 50%, only to slowly decay to results consistent with null-hypothesis [28]. This last behavior, 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 favor of imbalance. This in turn, tends to suggest that conflictual situations are less potent when distributed over many actors.

E. 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 [20, 21].

For the Monte Carlo implementation of Eq. (3), we sampled 20,000,000 graphs on 20 vertices from this net-

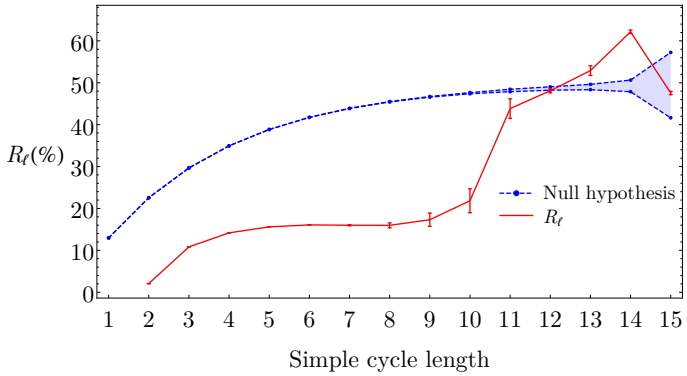

FIG. 5: Computed percentage of negatively signed simple cycles on the Epinions network for cycle length up to 15 (red line and error bars). The blue shaded region bordered by dashed blue lines shows the values of R_ℓ compatible with the null-hypothesis, as determined by Eqs. (6, 7).

work. We present the balance ratio R_ℓ up to $\ell = 20$ on Fig. (4) and give the full numerical results in Appendix C, Table III.

The balance ratio on this network exhibit a behavior similar to that observed on the WikiElection network: at first R_ℓ increases smoothly with $\ell \leq 10$. Then R_ℓ 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 $14 \leq \ell \leq 16$ 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.

F. 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. (3), we sampled 1,000,000,000 (one billion) graphs on 30 vertices from this network. We present the resulting balance ratio R_ℓ up to $\ell = 15$ on Fig. (5). We give the full numerical results up to $\ell = 20$ on Appendix C, Table IV. We were not able to compute the balance ratio $R_\ell^{\rm ob}$ using primitive orbits, owing to the very large size of this network.

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

V. CONCLUSION

A. Balance

By analyzing the simple cycles, we have shown that social networks are indeed strongly balanced. More precisely, the percentage of negatively signed simple cycles is greatly depressed as compared to an independent sampling scenario (the null hypothesis), typically up to lengths of circa 10. It is interesting that on the three large networks analysed here (Wikielections, Slashdot and Epinions), a rapid transition from balance (small R_ℓ values) to random ($R_\ell \sim R_\ell^{\rm null}$) 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, indicating that much of the imbalance is shifted to long simple cycles.

Our results tentatively suggest that the simplest model for the balance ratio R_ℓ on large (sparse) social networks is a step function, with the step located circa $\xi \sim 5$, that is $\ell \sim 10$. The value of R_ℓ for $\ell \leq 2\xi$ can be estimated from short cycles—e.g. triads or squares—while for $\ell \geq 2\xi$, R_ℓ should be around 50%. A better, more advanced model could perhaps use an error-function fit, which should however be justified on sociological grounds. In this respect, we hope that our results will generate further investigations in the study of interpersonal relationships.

B. Functions on simple cycles and simple paths

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 engine 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. (3) which is presented in [7]. 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 off i ($i = j$). Note, the matrix extension is *not* obtained upon just removing the trace from Eq. (3). This matrix formulation, in conjunction with a Monte Carlo approach as effected here, should permit the calculation of such functions as the path vertex-centrality of any vertex v on very large networks, which is defined as the total number of simple paths—not just the shortest ones—passing through vertex v.

Acknowledgments

P.-L. Giscard is grateful for the financial support provided by the Royal Commission for the Exhibition of 1851.

Appendix 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 II, III and IV.

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_ℓ 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_ℓ 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.

Appendix B: Proof of Proposition III.1

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

$$\zeta_{|G|}^{-1}(z) = \prod_{c\in[C]} \left(1 - z^{\ell(c)}\right) = \prod_{\ell} \left(1 - z^{\ell}\right)^{N_{\mathrm{ob};\,\ell}}, \quad \text{(B1)}$$

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

$$\zeta_{|G|}^{-1}(z) = \prod_{\ell} \left(1 - z^{\ell}\right)^{N_{\mathrm{ob};\,\ell}} = \det\left(\mathbf{I} - z|\mathbf{T}|\right). \quad \text{(B2)}$$

Now, on taking the logarithm on both sides we obtain

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

Equating the coefficient of z^ℓ on left and right hand sides then gives

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

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

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

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 $N^{+}_{\mathrm{ob};\,\ell} + N^{-}_{\mathrm{ob};\,\ell}$. □

Appendix C: Full numerical results

In this section we present the full numerical results obtained on the four networks mentioned earlier. We also give the exact balance ratios for the self-loops, backtracks and triangles, which are respectively given by

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

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

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

	Hexagons R_6	**Heptagons** R_7	**Octagons** R_8	**Nonagons** R_9	**Decagons** R_{10}
Exact	44.47%	48.64%	51.10%	52.33%	52.7%

	Hendecagon R_{11}	**Dodecagon** R_{12}	**Tridecagon** R_{13}	**Tetradecagon** R_{14}	**Pentadecagon** R_{15}
Exact	52.43%	51.77%	51.03%	50.46%	50.09%

	Hexadecagon R_{16}	**Heptadecagon** R_{17}	**Octadecagon** R_{18}	**Enneadecagon** R_{19}	**Icosagon** R_{20}
Exact	49.74%				

TABLE I: Exact balance ratios R_ℓ for $1 \leq \ell \leq 20$ obtained on the Gama network using Eq. (3).

	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%		

	Hexagons R_6	**Heptagons** R_7	**Octagons** R_8	**Nonagons** R_9	**Decagons** R_{10}
Monte Carlo	$23.48 \pm 0.03\%$	$26.22 \pm 0.03\%$	$28.84 \pm 0.19\%$	$31.2 \pm 0.31\%$	$32.82 \pm 0.54\%$

	Hendecagon R_{11}	**Dodecagon** R_{12}	**Tridecagon** R_{13}	**Tetradecagon** R_{14}	**Pentadecagon** R_{15}
Monte Carlo	$33.95 \pm 0.9\%$	$35.89 \pm 0.9\%$	$40.96 \pm 0.9\%$	$50.16 \pm 0.19\%$	$57.3 \pm 0.36\%$

	Hexadecagon R_{16}	**Heptadecagon** R_{17}	**Octadecagon** R_{18}	**Enneadecagon** R_{19}	**Icosagon** R_{20}
Monte Carlo	$54.38 \pm 3.41\%$	$53.85 \pm 4.57\%$			

TABLE II: Computed balance ratios R_ℓ for $1 \leq \ell \leq 20$ on the WikiElections network together with twice the standard deviation exhibited by the Monte Carlo results. We found no octadecagon, enneadecagon and icosagon on this network.

	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%		

	Hexagons R_6	**Heptagons** R_7	**Octagons** R_8	**Nonagons** R_9	**Decagons** R_{10}
Monte Carlo	$9.37 \pm 0.06\%$	$10.55 \pm 0.12\%$	$11.82 \pm 0.15\%$	$13.3 \pm 0.09\%$	$15.32 \pm 0.46\%$

	Hendecagon R_{11}	**Dodecagon** R_{12}	**Tridecagon** R_{13}	**Tetradecagon** R_{14}	**Pentadecagon** R_{15}
Monte Carlo	$16.5 \pm 0.8\%$	$21.3 \pm 2.6\%$	$32.7 \pm 1.5\%$	$52.9 \pm 3.6\%$	$58.4 \pm 2.6\%$

	Hexadecagon R_{16}	**Heptadecagon** R_{17}	**Octadecagon** R_{18}	**Enneadecagon** R_{19}	**Icosagon** R_{20}
Monte Carlo	$54.5 \pm 4.3\%$	$50.0 \pm 6.6\%$	$50.0 \pm 9.4\%$		

TABLE III: Computed balance ratios R_ℓ for $1 \leq \ell \leq 20$ on the Slashdot network together with twice the standard deviation exhibited by the Monte Carlo results. We found no enneadecagon and icosagon on this network.

	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%		

	Hexagons R_6	**Heptagons** R_7	**Octagons** R_8	**Nonagons** R_9	**Decagons** R_{10}
Monte Carlo	$16.07 \pm 0.02\%$	$15.99 \pm 0.12\%$	$15.97 \pm 0.58\%$	$17.30 \pm 1.58\%$	$21.83 \pm 2.84\%$

	Hendecagon R_{11}	**Dodecagon** R_{12}	**Tridecagon** R_{13}	**Tetradecagon** R_{14}	**Pentadecagon** R_{15}
Monte Carlo	$43.8 \pm 2.3\%$	$48.1 \pm 0.5\%$	$52.9 \pm 1.2\%$	$62.2 \pm 0.4\%$	$47.6 \pm 0.4\%$

	Hexadecagon R_{16}	**Heptadecagon** R_{17}	**Octadecagon** R_{18}	**Enneadecagon** R_{19}	**Icosagon** R_{20}
Monte Carlo		50%		44.4%	

TABLE IV: Computed balance ratios R_ℓ for $1 \leq \ell \leq 20$ on the Epinions network together with twice the standard deviation exhibited by the Monte Carlo results. We found no hexadecagon, octadecagon and icosagon 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. (7).

[1] F. Heider, The Journal of Psychology **21**, 107 (1946).
[2] D. Cartwright and F. Harary, Psychological Review **63**, 277 (1956).
[3] F. Harary, Behavioral Science **4**, 316 (1959).
[4] R. Z. Norman, Journal of Mathematical Psychology **9**, 66 (1972).
[5] F. Harary and J. A. Kabell, Mathematical Social Sciences **1**, 131 (1980).
[6] T. Antal, P. L. Krapivsky, and S. Red, Physica D: Nonlinear Phenomena **224**, 130 (2006).
[7] P.-L. Giscard and P. Rochet, arXiv:1606.00289 (2016).
[8] F. Harary and B. Manvel, Matematický časopis **21**, 55 (1971).
[9] E. Estrada and M. Benzi, Physical Review E **90**, 042802 (2014).
[10] R. B. Zajonc and E. Burnstein, Journal of Personality **33**, 570 (1965).
[11] A. Terras, *Zeta Functions of Graphs: A Stroll through the Garden* (Cambridge University Press, Cambridge, 2011), 1st ed.
[12] Y. Ihara, Journal of the Mathematical Society of Japan **18**, 219 (1966).
[13] H. M. Stark and A. A. Terras, Advances in Mathematics **121**, 124 (1996).
[14] G. Facchetti, G. Iacono, and C. Altafini, Proceedings of the National Academy of Sciences **108**, 20953 (2011).
[15] UCINET IV Datasets, *UCINET IV Datasets, Read Highland Tribes*, `http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm` (accessed in May 2016).
[16] Wikipedia adminship election data, *Wikipedia adminship election data*, `http://snap.stanford.edu/data/wiki-Elec.html` (accessed in May 2016).
[17] Slashdot social network, February 2009, *Slashdot social network, February 2009*, `http://snap.stanford.edu/data/soc-sign-Slashdot090221.html` (accessed in May 2016).
[18] Epinions social network, *Epinions social network*, `http://snap.stanford.edu/data/soc-sign-epinions.html` (accessed in May 2016).
[19] P. Hage, Anthropological Forum: A Journal of Social Anthropology and Comparative Sociology **3**, 280 (1973).
[20] C. A. Lampe, E. Johnston, and P. Resnick, Proceedings of Computer/Human Interaction 2007 Conference (Association for Computing Machinery, New York) pp. 1253–1262 (2007).
[21] J. Kunegis, S. Schmidt, A. S. Lommatzsch, J. Lerner, E. W. D. Luca, and S. Albayr, SIAM Conference on Data Mining (Society for Industrial and Applied Mathematics, Philadelphia) pp. 559–570 (2010).
[22] P.-L. Giscard and P. Rochet, arXiv:1601.01780 (2016).
[23] 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.
[24] One should keep in mind that since $P(z)$ determines the existence and number of Hamiltonian cycles on G, and unless $\mathrm{P} = \mathrm{NP}$, this exponential cost is, in principle, the best possible.
[25] The properties of Eq. (3) contrast it with other analytical formulas for counting simple cycles of small lengths [8]. First, these formulas work only on undirected graphs. Second, these formulas comprise a large number of terms, e.g. 17,476 terms when counting simple cycles of length 10. Finally, all of these terms involve the adjacency matrix of the full graph, so that getting e.g. R_{10}, is highly resource-consuming even on small graphs.
[26] 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 [22].
[27] 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 $\mathbf{A} - \mathrm{Diag}(\mathbf{A})$ in the calculations of $R_{\ell \geq 3}$.
[28] This effect becomes even more pronounced when R_ℓ is compared with the more accurate null model that takes the structural correlations between simple cycles into account.