

Detecting Core-Periphery Structures by Surprise

Jeroen van Lidth de Jeude, Guido Caldarelli, and Tiziano Squartini
IMT School for Advanced Studies, Piazza S.Francesco 19, 55100 Lucca - Italy
(Dated: April 22, 2019)

Detecting the presence of mesoscale structures in complex networks is of primary importance. This is especially true for financial networks, whose structural organization deeply affects their resilience to events like default cascades, shocks propagation, etc. Several methods have been proposed, so far, to detect *communities*, i.e. groups of nodes whose internal connectivity is significantly large. Communities, however do not represent the only kind of mesoscale structures characterizing real-world networks: other examples are provided by bow-tie structures, core-periphery structures and bipartite structures. Here we propose a novel method to detect statistically-significant *bimodular* structures, i.e. either bipartite or core-periphery ones. It is based on a modification of the *surprise*, recently proposed for detecting communities. Our variant allows for bimodular nodes partitions to be revealed, by letting links to be placed either 1) within the core part and between the core and the periphery parts or 2) between the layers of a bipartite network. From a technical point of view, this is achieved by employing a multinomial hypergeometric distribution instead of the traditional, binomial hypergeometric one; as in the latter case, this allows a p-value to be assigned to any given (bi)partition of the nodes. To illustrate the performance of our method, we report the results of its application to several real-world networks, including social, economic and financial ones.

PACS numbers: 89.75.Fb; 02.50.Tt; 89.65.Gh

INTRODUCTION

Detecting the presence of mesoscale structures in complex networks is of primary importance [1, 2]. This is especially true for financial networks, whose structural organization deeply affects their resilience to shocks propagation, node failures, etc. [3–6]. Several methods have been proposed, so far, to detect communities, i.e. groups of nodes whose “internal” connectivity is significantly large. Communities, however, do not represent the only kind of mesoscale structures characterizing real-world networks: other examples are provided by bow-tie, core-periphery and bipartite structures. In what follows, we will focus on the last two types of topological structures.

The intuitive notion of core-periphery network, as a configuration consisting of a densely-connected bunch of nodes (i.e. the core) and low-degree nodes preferentially connected to the core (i.e. the periphery ones) has been firstly formalized by Borgatti & Everett: in [7] a score function indicating the extent to which a given graph partition deviates from an ideal core-periphery configuration (where the core is *fully* connected and the peripheral nodes are *only* linked to the core ones) was defined. Several later works adopted the same approach [4, 5, 8], accompanying the error score with a significance level, computed on a properly-generated ensemble of networks (see [9] for a review on the topic). Detection of bipartitiveness has been approached similarly, by quantifying the deviation of an observed graph partition from the ideal bipartite configuration (where edges exist only between layers and not within them) [10, 11].

Conversely, in recent years the detection of mesoscale structures has been faced by adopting a bottom-up approach, i.e. by defining a benchmark model against which

to compare the actual network structure: in [12] the authors aim at identifying the most likely generative model that may have produced a given partition; in [13, 14] the authors compare the likelihood values of a Stochastic Block Model tuned to reproduce either a core-periphery or a bipartite structure; similarly, in [15] the authors adopt a Random Graph Model to find multiple core-periphery pairs in networks and in [16] the same authors employ the Configuration Model as a benchmark, showing that a single core-periphery structure can never be significant under it, seemingly confirming recent findings by the authors of the present paper [4, 17].

We contribute to this stream of research by proposing a novel method to detect statistically-significant bimodular structures (i.e. either bipartite or core-periphery ones). To this aim, we build upon the results of the papers [18–20] and on the very last comment that can be found in [21], by adopting a surprise-like score function. Our choice is dictated by the versatility of this kind of quantity that allows us to consider undirected as well as directed (binary) networks, a desirable feature that many of the aforementioned algorithms do not have.

The paper is organized as follows: the Methods section is devoted to illustrate the definition of our bimodular surprise; the results of its application to real-world networks are shown in the Results section and further discussed in the Discussion section where future perspectives are also presented.

METHODS

Let us first discuss the limitations of traditional surprise whenever employed to detect bimodular structures. In what follows we will implement the following definition of surprise [18–20]

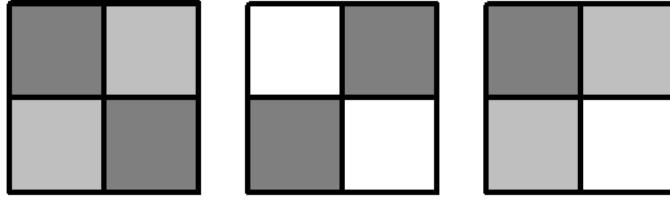


FIG. 1. Examples of mesoscale network structures: a traditional community structure is shown on the left, a purely bipartite network is shown in the middle and a core-periphery structure is shown on the right. White blocks represents subsets of nodes whose link density is zero, darker blocks represents subsets of nodes whose link density is higher.

$$S \equiv \sum_{i \geq l^*} \frac{\binom{V_{int}}{i} \binom{V - V_{int}}{L-i}}{\binom{V}{L}}, \quad (1)$$

the sum runs up to the value $i = \min\{L, V_{int}\}$, where V is the volume of the network, coinciding with the total number of nodes pairs (i.e. $V = \frac{N(N-1)}{2}$ in the undirected case and $V = N(N-1)$ in the directed case), V_{int} is the total number of intracluster pairs (i.e. the number of nodes pairs *within* the individuated communities), L is the total number of links and l^* is the observed number of intracluster links (i.e. *within* the individuated communities).

The hypergeometric distribution describes the probability of observing i successes in L draws (without replacement) from a finite population of size V that contains exactly V_{int} objects with the desired feature (in our case, being an intracluster pair), each draw being either a success or a failure: surprise is the p-value of such a distribution, testing the statistical significance of the observed partition against the null hypothesis that the intracluster link density $p_{int} = \frac{l^*}{V_{int}}$ is compatible with the density $p = \frac{L}{V}$ characterizing the (Directed) Random Graph Model.

The limitations of surprise

While traditional surprise S is suited for community detection, it suffers from several limitations whenever employed to detect bimodular mesoscale structures.

Bipartite networks. Let us first consider a purely bipartite, undirected network, as the one shown in fig. 1, whose first and second layer consist of N_1 and N_2 nodes respectively. Since we would like S to reveal two (empty) communities, we would be tempted to instantiate eq. 1 with the values $V = \frac{(N_1+N_2)(N_1+N_2-1)}{2}$, $V_{int} = \frac{N_1(N_1-1)}{2} + \frac{N_2(N_2-1)}{2}$ and $l^* = 0$; upon considering, however, that $L \leq V_{int}$, the explicit computation of S reveals that $S = 1$ (as follows from the Vandermonde identity). Since S is nothing else than a p-value, a significant partition is expected to satisfy $S \leq S_{th}$, with S_{th} usually chosen to attain the value 0.01 or 0.05. In our

case, however, the opposite result is obtained: the considered (bi)partition *cannot* be significant, independently from the actual number of connections characterizing the considered configuration. This example highlights one of the limitations of the definition provided in eq. 1.

Star-like networks. Let us now consider proper core-periphery networks: according to the intuitive definition provided in [7], such configurations are characterized by a densely-connected portion, i.e. the core (in the ideal case $c_c \simeq 1$) and a sparsely-connected portion, i.e. the periphery (in the ideal case $c_p \simeq 0$). The density of the intermediate portion is variable, although the chain of inequalities $c_p \leq c_{cp} \leq c_c$ is always assumed to hold. Let us consider a peculiar example of this kind of networks, i.e. an undirected configuration with a fully connected core plus a periphery of nodes, each of which is connected to just one core node. For the moment, let us suppose that the number of core nodes coincides with the number of periphery nodes and let us instantiate S on a partition that identifies each periphery node as a community on its own while considering the core as a traditional community (see fig. 2). If we consider a core portion of N_1 nodes and $N_2 = N_1$ peripheral nodes, we have $V = \frac{(N_1+N_2)(N_1+N_2-1)}{2}$, $V_{int} = \frac{N_1(N_1-1)}{2}$, $L = \frac{N_1(N_1-1)}{2} + N_1$ and $l^* = \frac{N_1(N_1-1)}{2}$. Notice that, in this case, only the addendum corresponding to the value $i = l^* = V_{int}$ survives, i.e.

$$S = \frac{\binom{N_1(3N_1-1)/2}{N_1}}{\binom{N_1(2N_1-1)}{N_1(N_1+1)/2}} \quad (2)$$

which is of the order of 10^{-2} for $N_1 = 3$ and rapidly decreases as N_1 grows (see fig. 2). Since $S < S_{th} = 0.05$, such a partition is recovered as significant. As confirmed by running the PACO algorithm [20], such a configuration - constituted by an unreasonably large number of single-nodes communities - is indeed recognized as the optimal one.

For the sake of comparison, let us calculate S for the “reasonable” partition identifying the core and the periphery as two separate communities: in this case, $V = \frac{(N_1+N_2)(N_1+N_2-1)}{2}$, $V_{int} = N_1(N_1-1)$, $L = \frac{N_1(N_1-1)}{2} + N_1$ and $l^* = \frac{N_1(N_1-1)}{2}$. As our explicit calculation reveals, such a partition can indeed be significant but it is not

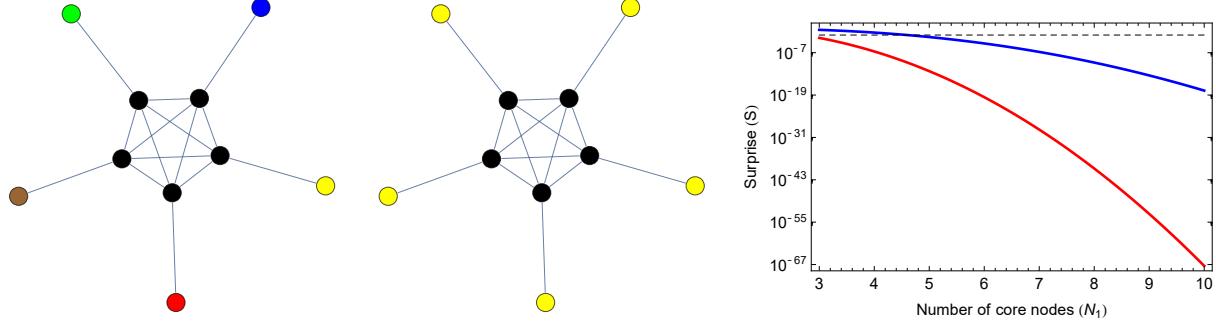


FIG. 2. Right panel: traditional surprise computed for the two partitions shown in the left and middle panels. The red line refers to the partition constituted by 6 communities (left panel), the blue line refers to the partition constituted by 2 communities (middle panel) and the black, dashed line corresponds to the value $S_{th} = 0.05$. As the number of core nodes is risen, both partitions become increasingly significant; the former, however, is always more significant than the latter. The network configuration shown in the left panel is, in fact, recognized as the optimal one, as further confirmed by running the PACO algorithm [20].

the optimal one (see also fig. 2).

k-star networks. Let us now generalize the star-like network model, by considering a graph with k peripherical nodes linked to each core node (see fig. 3). Instantiating S by considering each group of k leaves as a community on its own leads to

$$S = \sum_{i=l^*}^L \frac{\binom{V_{int}}{i} \binom{V-V_{int}}{L-i}}{\binom{V}{L}} \quad (3)$$

with $V = \frac{(N_1+kN_1)(N_1+kN_1-1)}{2}$, $V_{int} = \frac{N_1(N_1-1)}{2} + \frac{N_1k(k-1)}{2}$, $L = \frac{N_1(N_1-1)}{2} + kN_1$ and $l^* = \frac{N_1(N_1-1)}{2}$ (as long as $k \geq 2$, in fact, $L \leq V_{int}$). The expression defined by eq. 3 is significant only under certain conditions: in particular, *a)* for a given N_1 value, as k grows surprise becomes increasingly non-significant; *b)* for a given k value, as N_1 grows surprise becomes increasingly significant. Since the k nodes linked to each core node should be always considered as *non* constituting separate communities, irrespectively from the value of k , the findings above point out another detectability limit of surprise that, for certain values of the parameters, misinterprets the (planted) partition under analysis.

A bimodular surprise

The previous examples have shown that traditional surprise may suffer from some limitations whenever employed to detect bimodular structures. Here we address such an issue by introducing a variant of traditional surprise, designed to detect bimodular mesoscale structures.

Whenever community detection is carried out by maximizing the surprise, links are understood as belonging to *two* different categories, i.e. the *internal* ones (the ones *within* clusters) and the *external* ones (the ones *between* clusters). On the other hand, whenever one is interested in detecting bimodular structures (be they bipartite

or core-periphery), *three* different “species” of links are needed (e.g. core, core-periphery and periphery links). This is the reason why we need to consider the multinomial version of the surprise, whose definition reads

$$S_{\parallel} \equiv \sum_{i \geq l_c^*} \sum_{j \geq l_{cp}^*} \frac{\binom{V_c}{i} \binom{V_{cp}}{j} \binom{V-(V_c+V_{cp})}{L-(i+j)}}{\binom{V}{L}} \quad (4)$$

and that we will refer to as to the *bimodular surprise*. The presence of three different binomial coefficients allows three different kinds of links to be accounted for. From a technical point of view, S_{\parallel} is a p-value computed on a multivariate hypergeometric distribution describing the probability of $i+j$ successes in L draws (without replacement), from a finite population of size V that contains exactly V_c objects with a first specific feature *and* V_{cp} objects with a second specific feature, wherein each draw is either a success or a failure. Although i and j are respectively bounded by the values V_c and V_{cp} , analogously to the univariate case, $i+j \in [l_c^* + l_{cp}^*, \min\{L, V_c + V_{cp}\}]$.

The index c in eq. 4 labels the core part and the index cp labels the core-periphery part; whenever considering bipartite networks, the core-periphery portion will be assumed to indicate the inter-layer portion.

Bipartite networks. Let us now calculate S_{\parallel} for the bipartite case considered above, defined by the values of parameters $V_c = \frac{N_1(N_1-1)}{2}$ (here, the label c indicates the internal volume of one of the two layers), $V_{cp} = N_1N_2$, $l_c^* = 0$ and $l_{cp}^* = L$. The latter condition implies that only the addendum corresponding to $i = 0, j = l_{cp}^* = L$ survives; thus, our bimodular surprise reads

$$S_{\parallel} = \frac{\binom{V_{cp}}{l_{cp}^*}}{\binom{V}{l_{cp}^*}} = \frac{\binom{N_1N_2}{l_{cp}^*}}{\binom{(N_1+N_2)(N_1+N_2-1)/2}{l_{cp}^*}} \quad (5)$$

which *can* be significant, as it should be: in fact, a number of inter-layer links exists above which the observed

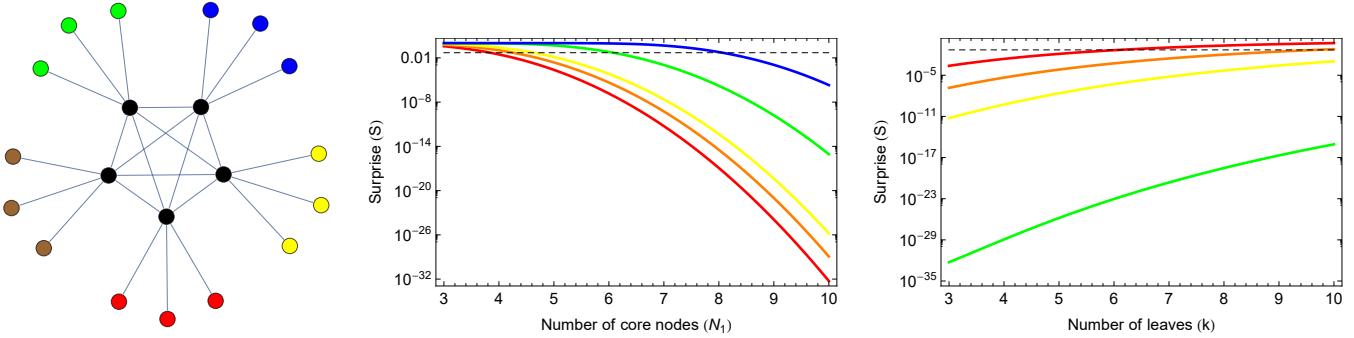


FIG. 3. Traditional surprise optimization on a k -star network would lead to identify each group of peripheral nodes as a community on its own, although the intracluster density is zero. More precisely, for a given number of leaves ($k = 3, 4, 5, 10, 20$ as indicated by the red, orange, yellow, green, blue line respectively - middle panel), as the number of core nodes rises, surprise is found to be increasingly significant. Consistently, for a given number of core nodes ($N_1 = 5, 6, 7, 10$ as indicated by the red, orange, yellow, green line respectively - right panel), surprise is increasingly non-significant as the number of leaves rises (the black, dashed line corresponds to the value $S_{th} = 0.05$). These findings point out the existence of a region of the parameter space where surprise misinterprets the planet partition.

bipartite structure is significantly denser than its random counterpart (see also fig. 4). Notice that eq. 5 can be directly employed to test the significance of any bipartite configuration with no intra-layer links, against the null hypothesis that such a configuration is compatible with the Random Graph Model: eq. 5 shows that, in the considered case, the computation of the searched p-value boils down to calculate the ratio between the number of bipartite networks with l_{cp}^* links and the number of generic configurations with the same number of connections.

Star-like networks. Let us now implement our bimodal surprise S_{\parallel} for star-like configurations. The core portion is identified with the clique of N_1 nodes: our parameters, thus, read $V_c = \frac{N_1(N_1-1)}{2}$ and $V_{cp} = N_1^2$. Since, however, $l_c^* = \frac{N_1(N_1-1)}{2}$, the (only) sum indexed by j reduces to the single addendum

$$S_{\parallel} = \frac{\binom{N_1^2}{N_1}}{\binom{N_1(2N_1-1)}{N_1(N_1+1)/2}} \quad (6)$$

which is $\simeq 10^{-2}$ for $N_1 = 3$ and decreases (the corresponding partition, thus, becomes more and more significant) as N_1 increases. Notice that the traditional surprise would identify a community structure - with each peripheral node counted as a community on its own - with a comparable significance (see also fig. 3): S_{\parallel} , however, is able to recover the ground-truth structure of the observed network.

k -star networks. Analogously, in the k -star case our parameters read $V = \frac{(N_1+kN_1)(N_1+kN_1-1)}{2}$, $V_c = \frac{N_1(N_1-1)}{2}$, $V_{cp} = kN_1^2$, $l_c^* = \frac{N_1(N_1-1)}{2}$ and $l_{cp}^* = kN_1$. Again, thus, the (only) sum indexed by j reduces to just one addendum, i.e.

$$S_{\parallel} = \frac{\binom{kN_1^2}{kN_1}}{\binom{(N_1+kN_1)(N_1+kN_1-1)/2}{N_1(N_1-1)/2+kN_1}} \quad (7)$$

whose behavior is shown in fig. 4: briefly speaking, both in case the number N_1 of core nodes rises, while keeping the number of leaves fixed, and the number k of leaves rises, while keeping the number of core nodes fixed, the bimodal surprise becomes increasingly significant, always recovering the ground-truth partition.

Asymptotic results

The presence of binomial coefficients in the definition of S_{\parallel} may cause its explicit computation to be demanding from a purely numerical point of view. This subsection is devoted to derive some asymptotic results, in order to speed up the computation of S_{\parallel} . Similar calculations for what concerns the traditional surprise have been carried out in [21].

Let us start by considering eq. 5. By Stirling expanding the binomial coefficients appearing in it, one obtains the expression

$$S_{\parallel} = \frac{\binom{V_{cp}}{l_{cp}^*}}{\binom{V}{l_{cp}^*}} \simeq \frac{p^{l_{cp}^*}(1-p)^{V-l_{cp}^*}}{p_{cp}^{l_{cp}^*}(1-p_{cp})^{V_{cp}-l_{cp}^*}} \quad (8)$$

having defined $p \equiv \frac{l_{cp}^*}{V}$ and $p_{cp} \equiv \frac{l_{cp}^*}{V_{cp}}$ (see the Appendix for the details of the calculations). The expression above makes it explicit that a given (bi)partition is statistically significant if its link density, p_{cp} , is large enough to let it be distinguishable from a typical configuration of the Random Graph Model, characterized by link density p .

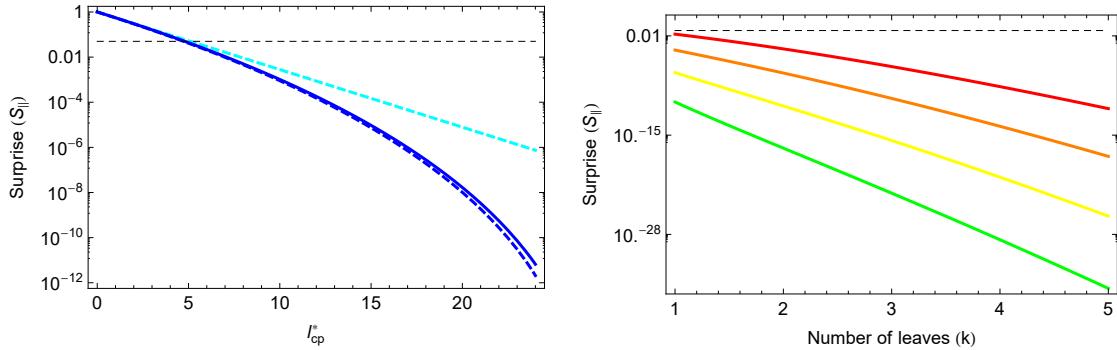


FIG. 4. Left panel: behavior of S_{\parallel} as a function of l_{cp}^* , for a bipartite network with $N_1 = N_2 = 5$. The blue, solid line corresponds to the (full) expression shown in eq. 5; the blue, dashed line corresponds to the asymptotic expression shown in eq. 8 and the cyan, dashed line corresponds to its sparse-case approximation. Right panel: behavior of S_{\parallel} for k -star network configurations (star-like networks are recovered as a particular case, when $k = 1$). For a given number of core nodes ($N_1 = 3, 4, 5, 6$ as indicated by the red, orange, yellow, green line respectively), surprise becomes increasingly significant, as the number of leaves rises. The black, dashed line corresponds to the value $S_{th} = 0.05$ in both cases.

In the sparse case, i.e. when $p \ll 1$ and $p_{cp} \ll 1$, eq. 8 reduces to $S_{\parallel} \simeq \left(\frac{p}{p_{cp}}\right)^{l_{cp}^*}$.

Let us now move to the core-periphery case and consider partitions satisfying the condition $l_c^* + l_{cp}^* = L < V_c + V_{cp}$: in this case, one can derive the result

$$S_{\parallel} = \frac{\binom{V_c}{l_c^*} \binom{V_{cp}}{l_{cp}^*}}{\binom{V}{L}} \simeq \frac{p^L (1-p)^{V-L}}{p_c^{l_c^*} (1-p_c)^{V_c-l_c^*} \cdot p_{cp}^{l_{cp}^*} (1-p_{cp})^{V_{cp}-l_{cp}^*}} \quad (9)$$

having defined $p \equiv \frac{L}{V} = \frac{l_c^* + l_{cp}^*}{V}$, $p_c \equiv \frac{l_c^*}{V_c}$ and $p_{cp} \equiv \frac{l_{cp}^*}{V_{cp}}$. Even if interpreting eq. 9 is less straightforward, it is, however, clear that the significance of the observed partition is a consequence of the interplay between the link density of the core and core-periphery regions (the link density of the periphery has been supposed to be zero - see the Appendix for the details of the calculations).

RESULTS

Let us now move to analyze some real-world systems: we will employ our novel definition of surprise to understand if the considered networks have a significant bimodular structure (i.e. either bipartite or core-periphery). To this aim, we will search for the (optimal) partition that minimizes S_{\parallel} by employing a modified version of the PACO algorithm [20] whose pseudocode is explicitly shown in Appendix and a Python version of which is freely available at https://github.com/jeroenvldj/bimodular_surprise. In what follows we will consider directed as well as undirected networks.

Social networks. Let us start our analysis by considering a number of undirected social networks (see fig. 5). As a first example, let us consider the Zachary Karate Club. Although the latter is commonly employed as a

benchmark for community detection, it is also characterized by a clear bimodular structure whose core nodes are represented by the masters, their close disciples and a fifth node “bridging” the two masters. Upon looking at the subgraphs constituted by the masters’ ego-networks, almost ideal (i.e. *à la Borgatti*) core-periphery networks are observable.

A similar comment can be done when considering the network of relationships among the characters of “Les Misérables”: the main characters (e.g. Valjean, Javert, Cosette, Marius) belong to the core, while the large number of secondary characters linked to them constitute the periphery of such a network (see, for example, the nodes linked to Valjean); intuitively, again, core nodes are very inter-connected while the link density of the periphery is very low. As for the Zachary Karate Club network, there seem to be (core) nodes bridging two dense core subsets.

Let us now consider the (connected component of the) NetSci co-authorship network [22]. A core-periphery structure is, again, recovered (although the core is not very dense) where core nodes represent senior scientists (e.g. Staneley, Barabasi, Watts, Kertesz) and periphery nodes represent younger colleagues, students, etc. It is interesting to observe that the senior scientists share relatively few direct connections, while being connected to a plethora of younger collaborators; even more so, the structure of the co-authorship network seems to reflect the structure of the underlying collaboration network, with each research group seemingly being quite separated from the others.

A fourth social network is the one showing the relationships between US political blogs [23]. Any two blogs are linked if one of the two references the other. As shown in fig. 6, a core of the most influential blogs (be they republican or democratic), surrounded by a periphery of loosely connected, less important blogs is clearly visible. Differently from the community structure that shows republican blogs and democratic blogs as belonging to dif-

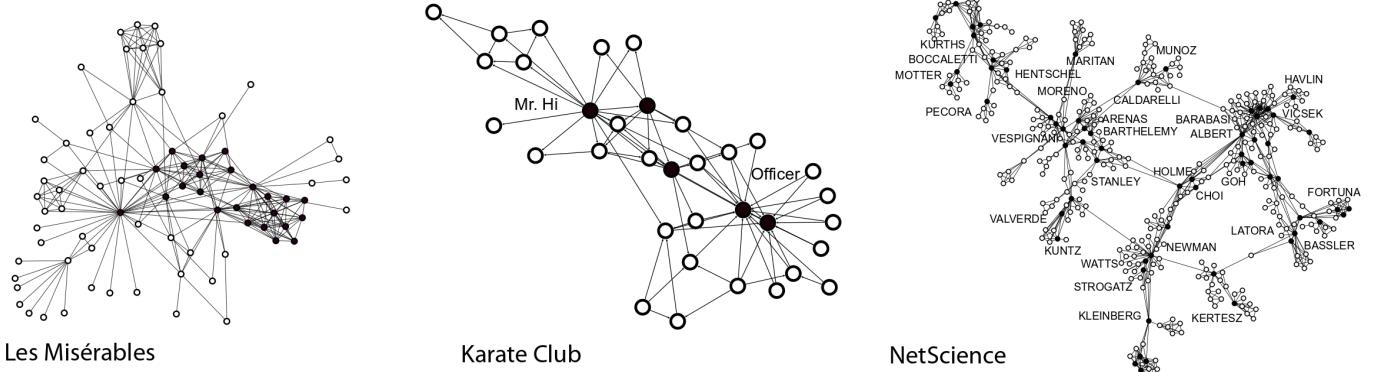


FIG. 5. Bimodular structure of three real-world social networks (core nodes are drawn in black and periphery nodes are drawn in white). Left panel: core-periphery structure of the network of relationships among “Les Miserables” characters; the main characters (e.g. Valjean, Javert, Cosette, Marius) belong to the core. Middle panel: core-periphery structure of the Zachary Karate Club; while the two masters (plus some close disciples) belong to the core, the remaining disciples create a periphery around them, shaping a configuration that is reminiscent of the Borgatti & Everett ideal structure [7]. Right panel: core-periphery structure of the NetSci co-authorship network; while the senior scientists belong to the core - although sharing few direct connections - younger colleagues/students belong to node-specific peripheries connected to the former ones.

ferent groups [24], our core-periphery structure highlights a different organizing principle, based on the blogs overall importance irrespectively from their political orientation. Interestingly enough, the bimodular surprise value indicates that the core-periphery structure is more significant than the traditional republicans VS democrats community structure.

Economic networks. Let us now consider an economic network, i.e. the directed representation of the World Trade Web (WTW) in the years 1950-2000: as usual, nodes are world countries and links are trade relationships (i.e. exports, imports) between them. Upon running our bimodular surprise optimization we find a clear core-periphery structure with the core including the richest countries and several developing nations and the periphery including some of the poorest nations (e.g. several African nations throughout our dataset - see also fig. 7 where only the years 1960, 1980 and 2000 are shown).

We also observe an interesting dynamics, causing the core size to rise (it represents the $\simeq 30\%$ of nodes in 1992 and the $\simeq 60\%$ of nodes in 2002) and progressively include countries previously belonging to the periphery. Such a dynamics - that can be interpreted as a signal of ongoing integration - confirms the results found in [17], where it was shown that the size of the WTW strongly connected component (SCC) increases with time as well. Although the SCC and the core portion of the World Trade Web do not perfectly overlap, many similarities between the two structures are indeed observable.

Financial networks. Let us now consider a financial network, i.e. e-MID, the electronic Italian Interbank Market. Here, we compare two different datasets: the first one collects the 2005-2010 interbank transactions during the so-called maintenance periods [25]; the second one collects interbank transactions on a daily basis from 1999 to 2012 [13, 14]. The main difference between

the two datasets lies in their level of aggregation: notice, in fact, that the first one basically collects data on a monthly basis.

Let us start by analyzing the first dataset. As fig. 8 shows, its structure undergoes an interesting evolution: after an initial period of two years, where a large periphery of loosely connected nodes ($\simeq 70\%$) exists, a transient period of one year (i.e. 2007) during which the percentage of nodes belonging to the core rises, is visible. Afterwards, an equilibrium situation seems to be re-established with the percentage of core and periphery nodes basically coinciding. Even if the total number of banks registered in the dataset steadily decreases after 2007, this doesn't seem to affect the type of banks belonging to the core and to the periphery, i.e. Italian and foreign banks, respectively.

Let us now move to the analysis of the second dataset. As fig. 9 shows, the analysis of the link density of the portions in which S_{\parallel} partitions the network reveals that, overall, a core-periphery structure seems to characterize the daily data better than a bipartite structure. This picture, however, seems to be less correct from 2008 on: as the last portion of the first panel of fig. 9 shows, a bipartite structures occur more often than a core-periphery structure during this period. Two snapshots of the network are also explicitly shown, illustrating the values of link density characterizing the different network portions.

Other kinds of networks. As a last example, let us consider the US airports network (see fig. 6). Examples of core airports are the ones of New York, Indianapolis, Salt Lake City, Seattle, etc. The periphery airports are preferentially attached to the core ones. This system shares interesting similarities with the NetSci co-authorship network: each core airport, in fact, seems to be surrounded by a quite large number of periphery airports, sharing few internal connections.

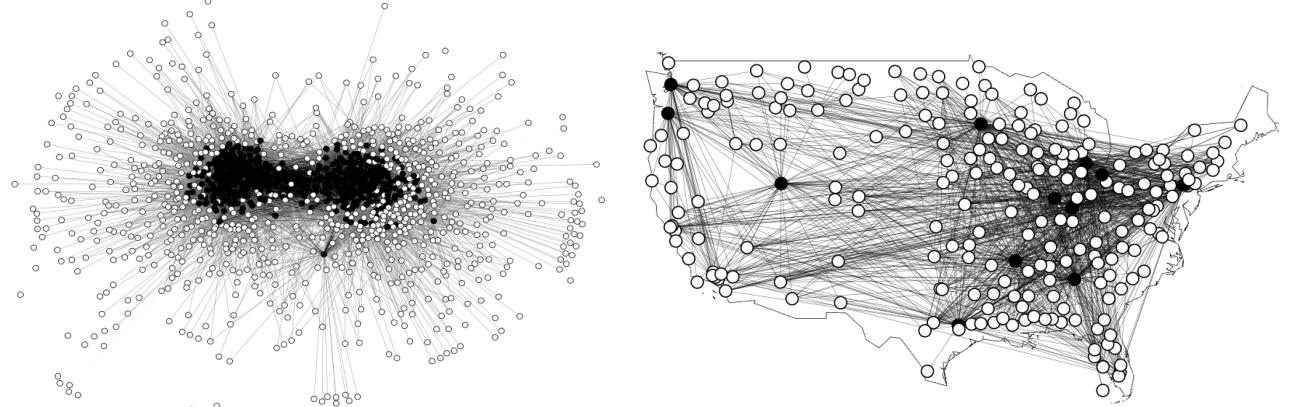


FIG. 6. Left panel: core-periphery structure of US political blogs [23]: a core of the most influential blogs (be they republican or democratic), surrounded by a periphery of loosely connected, less important blogs is clearly visible. Notice that blogs are grouped independently from their political orientation. Right panel: core-periphery structure of US airports. As for the NetSci co-authorship network, each core airport seems to be surrounded by a quite large number of periphery airports, sharing relatively few connections between themselves.

DISCUSSION

It is hard to underestimate the importance of the presence of bimodular mesoscale structures in real-world networks: while the authors in [26] show that the most robust topology against random failures is the core-periphery one, understanding the relationship between a given node systemicness and its coreness is of paramount importance in finance [6]. In the same field, a core-periphery structure is believed to reflect the “essential” function of banks: the core ones tie the periphery ones into a single market through their intermediation activity [3]. On the other hand, a bipartite structure would reflect the absence of intermediation, i.e. a market displaying preferential trading [13].

In this paper we have proposed a novel measure for bimodular mesoscale structures detection. To this aim, we have adopted a surprise-like score function, by considering the multivariate version of the quantity proposed in [20]. Employing this kind of quantities means implementing a bottom-up approach, i.e. letting the modular structure to be extrapolated from the data and not imposed *a priori* as in previous approaches [7, 14].

Most importantly, such a comparison is based on a properly-defined null model, allowing the significance of a given partition to be quantifiable via a p-value. As for the traditional surprise, the reference model is the (Directed) Random Graph Model that constrains the total number of observed connections, while randomizing everything else. The choice of employing such a benchmark is dictated by a number of recent results, pointing out that several mesoscale structures of interest (e.g. the core-periphery one, the bow-tie one, etc.) are actually compatible with - and hence undetectable under - a null model constraining the entire degree sequence(s) [4, 16].

While solving the problem of consistently comparing an observed structure with a “random” model of it, our

approach also solves a second drawback affecting the methods in [3, 7] and pointed out in [16]: ideal structures as the ones searched by algorithms *à la Borgatti* are very reliant on the nodes degree, with the core often composed of just the nodes with the largest number of neighbors. This is not necessarily true when a benchmark is adopted for comparison [12]: as previously discussed, the significance of a given partition detected by surprise results from the interplay between the link density values of the different network areas.

This also sheds light on the relationship between apparently conflicting structures co-existing within the same network configuration: generally speaking, traditional and bimodular surprise optimization should be considered complementary - rather than mutually exclusive - steps of a more general analysis. As the example of the US political blogs confirms, it is indeed possible that a community structure co-exists with a core-periphery structure; a second, less trivial, example is provided by the World Trade Web, whose community structure has been studied in [27] but whose significance has, then, been questioned [28].

As a last comment, we would like to stress that the two approaches to mesoscale structures detection that have been proposed so far - comparing an observed structure with a benchmark [15, 16] and searching for the model best fitting a given partition [12–14, 24] - can be supposed to be complementary, since a non-significant structure under a given benchmark is surely more compatible with it. Employing a benchmark, however, provides an advantage, i.e. making the statistical significance of a given structure explicit - something that remains “implicit” when employing the fitting procedure. In other words, searching for the best fit may push one to enrich a model with an increasing amount of information whose relevance cannot be easily clarified. Such a problem seems to affect all likelihood-based algorithms unless

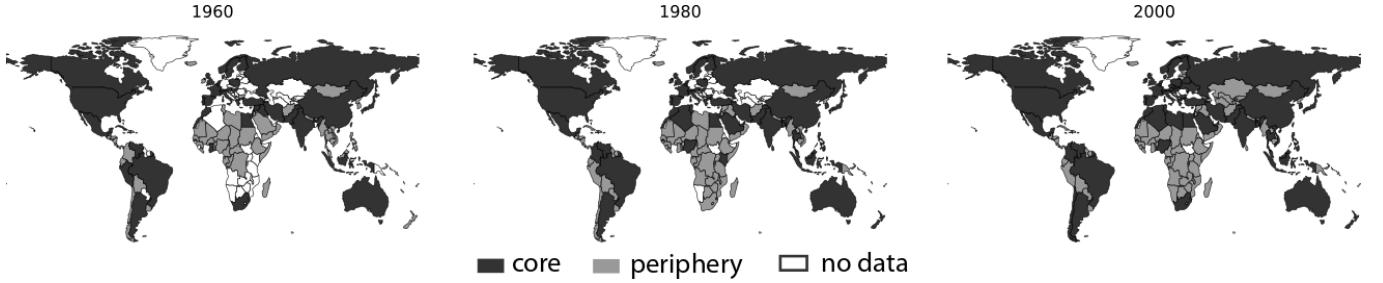


FIG. 7. Core-periphery structure of the World Trade Web (black: core nodes; gray: periphery nodes). Loosely speaking, while the richest and several developing countries are found to belong to the core, the poorest nations belong to the periphery (e.g. several African nations, throughout our dataset). Notice that core size increases with time: apparently, thus, the system becomes increasingly integrated, confirming a result found in [17], where it was shown that the size of the WTW strongly connected component increases with time as well.

a more refined criterion to judge the goodness of a fit is employed: solutions like the one of adopting criteria like the Akaike Information Criterion *et similia* have been proposed [29].

The present work calls for a generalization to *weighted* mesoscale structures detection, a field where relatively little has been done so far [30, 31].

APPENDIX

The computation of binomial coefficients for large graphs can quickly become numerically demanding. In order to simplify the calculations of our bimodular surprise, let us proceed by steps. First, let us Stirling approximating the binomial coefficients:

$$\binom{V_c}{i} \simeq \left[(p_c)^i (1-p_c)^{V_c-i} \right]^{-1}, \quad (10)$$

$$\binom{V_{cp}}{j} \simeq \left[(p_{cp})^j (1-p_{cp})^{V_{cp}-j} \right]^{-1}, \quad (11)$$

$$\binom{V_p}{L-(i+j)} \simeq \left[(p_p)^{L-(i+j)} (1-p_p)^{V_p-(L-(i+j))} \right]^{-1}, \quad (12)$$

$$\binom{V}{L} \simeq \left[p^L (1-p)^{V-L} \right]^{-1} \quad (13)$$

having defined $V_p \equiv V - (V_c + V_{cp})$, $p \equiv \frac{L}{V}$, $p_c \equiv \frac{i}{V_c}$, $p_{cp} \equiv \frac{j}{V_{cp}}$, $p_p \equiv \frac{L-(i+j)}{V_p}$. As a second step, let us substitute the

expressions above into eq. 4:

$$S_{\parallel} \simeq \sum_{i \geq l_c^*} \sum_{j \geq l_{cp}^*} \left(\frac{p}{p_p} \right)^L \left(\frac{1-p}{1-p_p} \right)^{V-L} \left(\frac{p_p}{p_c} \right)^i \left(\frac{1-p_p}{1-p_c} \right)^{V_c-i} \left(\frac{p_p}{p_{cp}} \right)^j \left(\frac{1-p_p}{1-p_{cp}} \right)^{V_{cp}-j}; \quad (14)$$

in order to obtain a more explicit expression, let us limit ourselves to consider the leading term of the summation

in eq. 14 that is readily obtained upon substituting i with l_c^* and j with l_{cp}^* :

$$S_{\parallel} \simeq \left(\frac{p}{p_p} \right)^L \left(\frac{1-p}{1-p_p} \right)^{V-L} \left(\frac{p_p}{p_c} \right)^{l_c^*} \left(\frac{1-p_p}{1-p_c} \right)^{V_c-l_c^*} \left(\frac{p_p}{p_{cp}} \right)^{l_{cp}^*} \left(\frac{1-p_p}{1-p_{cp}} \right)^{V_{cp}-l_{cp}^*} \quad (15)$$

where, now, $p_c \equiv \frac{l_c^*}{V_c}$, $p_{cp} \equiv \frac{l_{cp}^*}{V_{cp}}$, $p_p \equiv \frac{L-(l_c^*+l_{cp}^*)}{V_p}$. Notice

that eq. 15 can be employed to detect both bipartite and

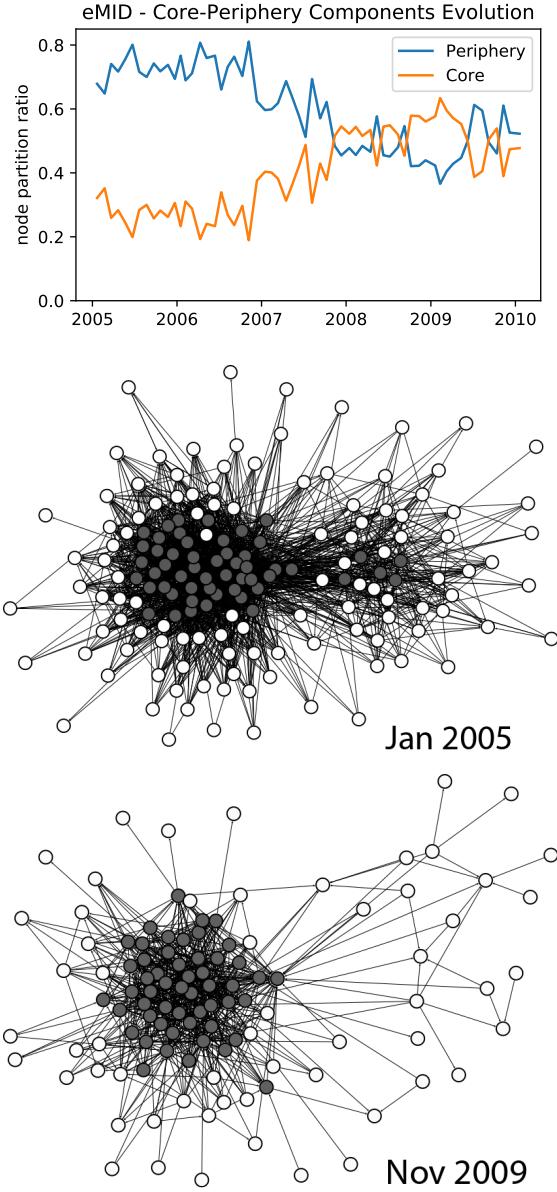


FIG. 8. Core-periphery structure of e-MID maintenance periods (gray: core nodes; white: periphery nodes). After an initial period of two years characterized by an approximately constant value of the core and periphery size, a structural change takes place in 2007 and the percentage of nodes belonging to the core steadily rises until 2008. Afterwards, an equilibrium seems to be re-established. This may be due to a decrease in the total number of nodes which, however, does not affect the type of banks belonging to the core (italian banks) and to the periphery (foreign banks). Networks are directed but we have omitted the link directionality for the sake of readability.

core-periphery structures, upon identifying the core and the periphery portions as the network portions *within* layers. Eq. 15 already makes intuitively clear that our bimodular surprise is likely to be significant either when $p_c \simeq p_p$ but $p_{cp} \gg p_c \simeq p_p$ (i.e. in the case of bipartite

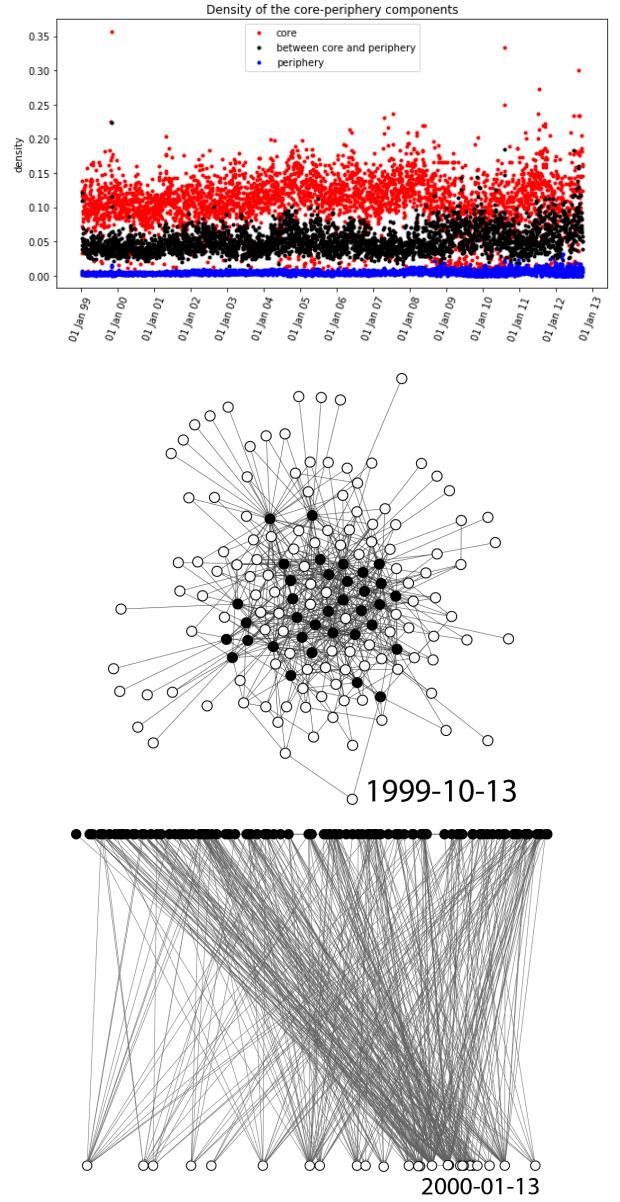


FIG. 9. Mesoscale structure of e-MID daily data. Although for the vast majority of snapshots a core-periphery structure seems to better represent the e-MID network, the number of times a bipartite structure is observed increases after 2008. Middle and bottom panels explicitly show two different snapshots of e-MID: the first one is characterized by the chain of inequalities $c_p < c_{cp} < c_c$; the second one, instead, shows a configuration for which the values $c_p \simeq c_c < c_{cp}$ are observed, indicating the presence of a bipartite structure (when referring to bipartite structures, the label cp is assumed to indicate the inter-layer portion). Networks are directed but we have omitted the link directionality for the sake of readability.

networks - notice that eq. 8 is recovered when $L = l_{cp}^*$ or when $p_c \gg p_p$ and $p_{cp} \gg p_p$ (i.e. in the core-periphery case - notice that eq. 9 is recovered when $L = l_c^* + l_{cp}^* < V_c + V_{cp}$).

A numerical check of the validity of the proposed ap-

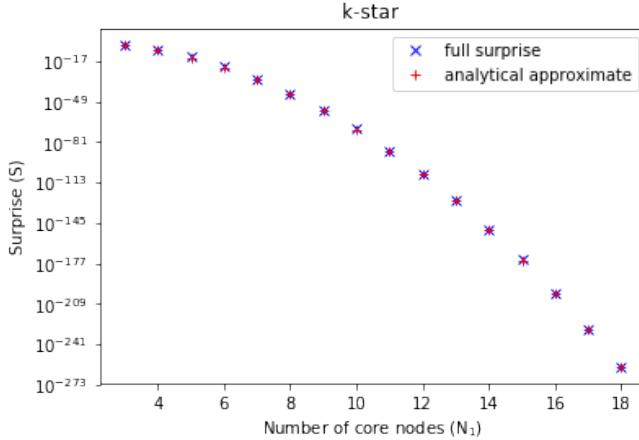


FIG. 10. Numerical check of the validity of the approximation shown in eq. 15, computed for k-star networks and compared to the full expression in eq. 4.

proximation is shown in fig. 10, where it is computed for k-star networks and compared to the full expression in eq. 4. We explicitly notice, however, that the approximation shown in eq. 15 is recommended for analysing small networks; when large networks are considered, the full expression in eq. 4 is, instead, recommended.

Numerical surprise optimization: a modified PACO algorithm

In what follows we show a modified version of the PACO (PArtitioning Cost Optimization) algorithm pseudocode [20] by running which the partition that minimizes surprise can be found. The PACO algorithm implements an approach that is heuristic in nature since an exhaustive search of all possible partitions is not feasible when dealing with large graphs.

The idea is that of assigning every node to either one of two subsets - interpretable as the core and the periphery or the layers of a bipartite graph - by running a greedy process that takes as input pairs of nodes connected by an edge and evaluates whether those two nodes should belong to the same subset or not: the choice that minimizes the surprise is the one that is actually implemented. In order to speed up the calculations, the original PACO algorithm takes edges that are previously sorted according to their decreasing value of Jaccard index. Since the latter quantifies the fraction of common neighbours of the two connected nodes, nodes pairs with larger Jaccard index are also the ones most likely to be assigned to the same subset (e.g. a community).

Since for pure bipartite graphs the Jaccard index - as defined above - is zero for all edges (nodes connected by an edge always lie on different layers) we need to modify the score according to which we sort the edges. In our modified version of the PACO algorithm we sort links

according to the number of z-motifs they belong to, the latter being defined as $z_{i\alpha} = \sum_{\beta,j} a_{i\beta} a_{i\alpha} a_{\alpha j}$: in other words, we evaluate the number of times a generic link is the “middle” one of a path whose length is 3. As with the original PACO algorithm, we progressively consider all edges, sorted as described above, evaluating whether the linked pairs should belong to the same subset or not.

As a final step, we consider a number of random reassignments of nodes with the aim of preventing the possibility of getting stuck in a local minimum (a random move consists of selecting 3 random nodes belonging to the same group and evaluating if assigning them to different subsets would further minimize surprise).

The algorithm described above performs quite well in finding the global minimum of surprise on a range of different configurations we have tested. When considering low-density bipartite graphs, however, the algorithm does not always succeed in reaching the global minimum.

```

1: function CALCULATEANDUPDATESURPRISE( $C, C'$ )
2:    $S \leftarrow \text{calculateSurprise}(C)$ 
3:    $S' \leftarrow \text{calculateSurprise}(C')$ 
4:   if  $S' < S$  then
5:      $C \leftarrow C'$ 
6:      $S \leftarrow S'$ 
7:   end if
8:   return  $C$ 
9: end function
10:
11:  $C \leftarrow$  array of length  $N$  randomly initialized with
    binary entries (0 or 1);
12:  $E \leftarrow$  sorted edges in decreasing order;
13: for edge  $(u, v) \in E$  do
14:    $C' \leftarrow C$ 
15:   if  $C'[u] \neq C[v]$  then
16:      $C'[u] \leftarrow C[v]$ 
17:      $C \leftarrow \text{CALCULATEANDUPDATESURPRISE}(C, C')$ 
18:   else
19:      $C'[u] \leftarrow 1 - C[v]$ 
20:      $C \leftarrow \text{CALCULATEANDUPDATESURPRISE}(C, C')$ 
21:   end if
22:    $\Rightarrow$  randomly switch node membership for  $n = 3$ 
    nodes in the same partition and accept move if  $S_{||}$ 
    decreases;
23: end for
24:  $\Rightarrow$  repeat several times the for-loop to improve the
    chance of finding the optimal partition.

```

ACKNOWLEDGEMENTS

This work was supported by the EU projects CoeGSS (grant num. 676547), DOLFINS (grant num. 640772), MULTIPLEX (grant num. 317532), Openmaker (grant num. 687941), SoBigData (grant num. 654024).

AUTHORS CONTRIBUTIONS

JLJ and TS developed the method. JLJ performed the analysis. JLJ, GC and TS wrote the manuscript. All

authors reviewed and approved the manuscript.

ADDITIONAL INFORMATION

The authors declare no competing financial interests.

- [1] S. Fortunato, D. Hric, Community detection in networks: a user guide, *Phys. Rep.* **659**, 1-44 (2016).
- [2] B. S. Khan, M. A. Niazi, Network community detection: a review and visual survey, *arXiv:1708.00977* (2017).
- [3] B. Craig, G. von Peter, Interbank tiering and money center banks, *J. Finan. Intermediation* **23** (3), 322-347 (2014).
- [4] D. in 't Veld, I. van Lelyveld, Finding the core: Network structure in interbank markets, *J. Bank. Finance* **49**, 27-40 (2014).
- [5] D. Fricke, T. Lux, Core-Periphery Structure in the Overnight Money Market: Evidence from the e-MID Trading Platform, *Computational Economics* **3** (45), 359-395 (2014).
- [6] D. T. Luu, M. Napoletano, P. Barucca, S. Battiston, Collateral Unchained: Rehypothecation networks, concentration and systemic effects, *Sciences Po OFCE working paper n. 07, 2018/01/31* (2018).
- [7] S. P. Borgatti, M. G. Everett, Models of core/periphery structures, *Soc. Networks* **21** (4), 375-395 (2000).
- [8] J. P. Boyd, W. J. Fitzgerald, R. J. Beck, Computing core/periphery structures and permutation tests for social relations data, *Soc. Networks* **28** (2) 165-178 (2006).
- [9] P. Csermely, A. London, L.-Y. Wu, B. Uzzi, Structure and dynamics of core-periphery networks, *J. Comp. Nets.* **1**, 93-123 (2013).
- [10] P. Holme, F. Liljeros, C. R. Edling, B. J. Kim, Network bipartivity, *Phys. Rev. E* **68** (5) (2003).
- [11] E. Estrada, J. A. Rodríguez-Velázquez, Spectral measures of bipartivity in complex networks, *Phys. Rev. E* **72** (4), 1-16 (2005).
- [12] X. Zhang, T. Martin, M. E. J. Newman, Identification of core-periphery structure in networks, *Phys. Rev. E* **91**, 032803 (2014).
- [13] P. Barucca, F. Lillo, Disentangling bipartite and core-periphery structure in financial networks, *arXiv:1511.08830v1* (2015).
- [14] P. Barucca, F. Lillo, The organization of the interbank network and how ECB unconventional measures affected the e-MID overnight market, *Comput. Manag. Sci.* **15** (1), 33-53 (2018).
- [15] S. Kojaku, N. Masuda, Finding multiple core-periphery pairs in networks, *Phys. Rev. E* **96** (052313) (2017).
- [16] S. Kojaku, N. Masuda, Core-periphery structure requires something else in the network, *New J. Physics* **4** (20), 359-395 (2018).
- [17] J. van Lidth de Jeude, R. Di Clemente, G. Caldarelli, F. Saracco, T. Squartini, Reconstructing mesoscale network structures, *arXiv:1805.06005* (2018).
- [18] R. Aldecoa, I. Marin, Surprise maximization reveals the community structure of complex networks, *Sci. Rep.* **3** (1060) (2013).
- [19] R. Aldecoa, I. Marin, Exploring the limits of community detection strategies in complex networks, *Sci. Rep.* **3** (2216) (2013).
- [20] C. Nicolini, A. Bifone, Modular structure of brain functional networks: breaking the resolution limit by Surprise, *Sci. Rep.* **6** (19250) (2016).
- [21] V. A. Traag, R. Aldecoa, J.-C. Delvenne, Detecting communities using asymptotical Surprise, *Phys. Rev. E* **92**, 022816 (2015).
- [22] <http://vlado.fmf.uni-lj.si/pub/networks/data/collab/netscience.htm>
- [23] L. A. Adamic, N. Glance, The Political Blogosphere and the 2004 U.S. Election: Divided They Blog, *Proceedings of the 3rd International Workshop on Link Discovery*, 36-43, ACM, New York (2005).
- [24] B. Karrer, M. E. J. Newman, Stochastic blockmodels and community structure in networks, *Phys. Rev. E* **83**, 016107 (2011).
- [25] V. Hatzopoulos, G. Iori, R. N. Mantegna, S. Micciché, M. Tumminello, Quantifying preferential trading in the e-MID interbank market, *Quantitative Finance* **15** (4), 693-710 (2015).
- [26] T. P. Peixoto, S. Bornholdt, Evolution of robust network topologies: emergence of central backbones, *Phys. Rev. Lett.* **109**, 118703 (2012).
- [27] M. Barigozzi, G. Fagiolo, G. Mangioni, Identifying the community structure of the international-trade multi-network, *Physica A* **390** (11), 2051-2066 (2011).
- [28] C. Piccardi, L. Tajoli, Existence and significance of communities in the World Trade Web, *Phys. Rev. E* **85**, 066119 (2012).
- [29] K. P. Burnham, D. R. Anderson, Model selection and multi-model inference: a practical information-theoretic approach, Springer, New York (2002).
- [30] C. Nicolini, C. Bordier, A. Bifone, Community detection in weighted brain connectivity networks beyond the resolution limit, *Neuroimage* **1** (146), 28-29 (2016).
- [31] G. Fagiolo, J. Reyes, S. Schiavo, On the topological properties of the world trade web: a weighted network analysis, *Physica A* **387** (15), 3868-3873 (2008).