The nested assembly of collective attention in online social systems

Javier Borge-Holthoefer,1,2* Raquel A. Baños,2,3 Carlos Gracia-Lázaro,2,3 Yamir Moreno2,3,4*

1Qatar Computing Research Institute, Doha, Qatar
2Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza,
50018 Zaragoza, Spain
3Department of Theoretical Physics, University of Zaragoza, Zaragoza 50009, Spain
4ISI Foundation, Turin, Italy

*To whom correspondence should be addressed: jborge@qf.org.qa, yamir.moreno@gmail.com

Online social networks have changed the way in which we communicate. These networks are characterized by a competitive information flow and a dynamic topology, where the success of a given topic or meme depends on many factors, most of which are unknown. Likely, given the many sources of information to which a typical individual is exposed, the economy of attention rules the system dynamics. Here we show, using microblogging data, that competition is minimized through consensus and that collective attention and successful topical assembly are characterized by a nested structure of the bipartite network made up by users and memes. Our results indicate that social phenomena could emerge as a result of a topological transition that minimizes modularity while maximizing the nestedness of the system, and that online social networks are comparable to an ecosystem, where generalists and specialists share resources.

Current state-of-the-art approaches to collective attention and information life-cycle[1–4] are aimed to prove how topics arise and fade in well-defined dynamical classes, with either the accent placed on collective human temporal dynamics or on semiotic dynamics[5], where in the latter case the temporal dimension is overlooked to provide cumulated evidence of how self-organization emerges in user-driven information systems[6]. Both frameworks, however, neglect user-meme (tags, words, hashtags) interactions because the originally rich bipartite representation of data is collapsed onto one of the dimensions of the problem. Therefore,
they only provide a partial understanding of how attention consensus around a certain topic builds up.

On the other hand, recent works\cite{7, 8} have highlighted the competitive nature of information systems, in which humans \textit{and} memes strive for scarce resources—visibility and attention, respectively. Here, we apply a different approach that has been proved to be very successful in biological systems to show that such a competitive environment facilitates the emergence of modular-to-nested architectures. Specifically, we consider information-driven systems such as online social platforms as mutualistic networks\cite{9, 11}, in which users and memes (the nodes of the bipartite graph) compete for resources within their respective classes. Our methodology allows observing modular and nested structures for different types of topics from large, public collections of on-line time-stamped microblogging data. More importantly, we show that competition between users and topics is mainly manifested in a modular structure, which leads as time goes on to a nested pattern where such competition is reduced and a consensus around the use of a certain set of memes is reached.

To validate our results, we use data corresponding to well-known phenomena and events. First, we build a series of time-stamped bipartite networks from the Twitter stream, filtered by a set of 70 keywords from a well-studied civil uprise that took place in Spain in 2011\cite{12}—see SOM for a detailed description of the procedure and Figure ??a for a snapshot of the resulting network. These data contain all the necessary information to build time-resolved bipartite networks—who said what, and when. The time-dependent bipartite graphs are built up as follows, see also Fig. ??b. First, time windows are set to a fixed, but arbitrary, $w = t_2 - t_1$ width. We then choose $n$ active users and $m$ memes (hashtags) within that time interval. This bipartite network is encoded in an $n \times m$ rectangular binary matrix, $M_t$, where $t$ indicates the origin of the time window $w$ and $M_{u,h} = 1$ if user $u$ mentioned the hashtag $h$ within the period spanning from $t_1$ to $t_2$ and zero otherwise. This procedure allows generating bipartite networks as time advances by using a sliding-window scheme to evaluate the evolution of the system, such that a window at time $t$ has a $\delta w$ overlap with that at time $t - 1$, see the \textit{SOM} for more details and an extended analysis.

Once the networks associated to the 15M social movement at different times are assembled, we proceed to analyze their structure focusing on two topological characteristics. As the interest is in inspecting whether groups of individuals using the same memes build up, we look for the optimal modular partition of the nodes through a community detection
analysis\cite{13}. To this purpose, among the many existent methods, we implement the fast algorithm heuristics by Newman\cite{14}, which is aimed at finding the partition that maximizes the so-called modularity

\[ Q = \frac{1}{2L} \sum_{i,j} \left( a_{ij} - \frac{k_i k_j}{2L} \right) \delta(C_i, C_j), \]

where \( L \) is the number of interactions (links) in the network, \( a_{ij} \) is 1 if there is a link from \( i \) to \( j \) (0 otherwise), \( k_i \) is the degree of node \( i \) and \( \delta \) is the Kronecker delta function, which takes the value 1 if nodes \( i \) and \( j \) are in the same community, and 0 otherwise. The SOM shows that results obtained with this method are robust when compared to other heuristics.

Secondly, we inspect whether nested structures arise in the system. A nested pattern corresponds to a situation in which interactions are arranged such that a few memes (individuals), henceforth referred to as generalists, are linked to many individuals (memes), whereas specialists are those memes (individuals) that interact with proper subsets of the individuals (memes) with which the generalists are linked to. Here, we evaluate the nestedness following the definition given by Bell\cite{15} and further developed in Staniczenko et al.\cite{13}, who showed that it is given by the maximum eigenvalue of the adjacency matrix, \( M_t \), of the networks. In the SOM we show that this definition is equivalent to that used in \cite{11} and perform a robustness analysis.

Figure ?? shows results of the application of the previous analyses for the 15M dataset and \( w = 3 \) days. At the beginning of the observation time (April 25th, 2011), the network exhibits the maximal modularity and the lowest nestedness value. This means that before the general onset of collective attention around the 15M activity, the (proto) topic is a set of compartmentalized groups (Fig. ??) which hardly interact with the rest of the system. At the same time, the structure of the network is not nested (Fig. ??). This picture however changes as the movement gains momentum and consensus arises. Indeed, the initially sparse network grows and as time progresses, we observe the opposite trend, i.e., the nestedness increases as the modularity decreases in an almost perfect anti-correlated pattern. After the crossover, the architecture of the networks is radically different.

The compelling evidence of nested patterns provides a parsimonious explanation of how large amounts of activity can coexist with natural constraints to attention and memory. The user-meme network self-organizes towards a nested structure minimizing competition and facilitating the coexistence of individual participants\cite{11}. Even when the network is predominantly modular, nestedness appears to have significant values well beyond random counterparts, which already indicates the existence of an incipient consensus around certain
topics. Moreover, the unraveled structural change from a configuration that maximizes modularity to a highly nested architecture allows interpreting the evolution of the Spanish mobilization episodes as a build-up effort from segregation (scattered activists acting locally) to coordination (a global movement with a well-defined and shared main message). Thus, our findings illustrate yet another example of commonalities between ecological and human\cite{17, 18} (even primate\cite{19}) systems.

As seen from Fig. ??, once the nested patterns begin to dominate the network structure –some days before the movement fully develops–, the nestedness remains at high levels for some time. This makes it possible to consistently track the set of users and memes that accumulate many interactions (generalists) and inspect whether these sets are time independent. To this end, we identify which nodes and which memes form the core of the networks at different times \cite{20}, see SOM for further details. Figure ?? compares the cores found at each snapshot to the ”gold standard” core, i.e., the one extracted when the nestedness is maximal. The comparison trivially renders total resemblance on such a moment, but the interest lies in the surroundings of it. Notably, for $w = 12h$ (top panel) there is a high turnover in users who occupy the generalist core, and yet the network architecture remains quite constant (see Fig. ??). Instead, hashtags have a much more stable core – around 20% of the core is shared during all the observation window, and values above 50% are reached after the movement onset and beyond. These results suggest that it is the topic, rather than the existence of generalist individuals, that guarantees consensus survival at the core of the system.

The previous finding qualitatively agrees with results reported for biological, temporally-resolved records\cite{21, 22} –which admittedly obey heavier observational constraints. Remarkably, high temporal volatility in users compositions coupled with low variation in the network structure imply that user-meme interaction networks might be resilient against disturbance, be it given by the removal of some memes or the introduction of new ones. Higher $w$ values yield similar results for most of the observation window, although cores show higher stability (for both hashtags and users) around the protest onset dates.

To show that our results are general and not specific to social phenomena of the kind of the 15M movement, we have analyzed an unfiltered dataset of Twitter traffic corresponding to geolocalized tweets in United Kingdom (see SOM for more details). As before, bipartite user-hashtag networks are built, but now we chose the subset comprising the top 1024 most-
active users and, independently, the subset of 1024 most-used hashtags. Note that such independent sampling implies that the corresponding adjacency matrix can be empty –the most active users might not use the most popular hashtags. Figure ?? shows the results obtained for this dataset. In the top panels, strongly fluctuating patterns are observed for both the modularity and the nestedness. Indeed, this does not resemble the more persistent, smoothly-developed 15M movement. On the contrary, most online topics that succeed in getting collective attention do not demand for days to brew and emerge, but they arise and decay at very fast time scales [4]. Despite the fact that in this case the temporal scale is much smaller, the same phenomenology is observed as illustrated in the bottom panels: collective attention around this topic –the XLVIII Super Bowl that started on February 3rd, 2014 at 12:30AM CET–, is reached when the network is maximally nested and minimally modular.

In summary, the analysis of two different datasets have shown how scattered pieces of information evolve towards a built-up nested information ecosystem. Therefore, whereas competition usually restricts the coexistence of different topics, a nested architecture allows for the existence of generalist memes that represent consensus around a certain topic. Our findings hint to the fact that user-meme systems engage mutualistic interactions [9], rendering optimal networked structures for the coexistence of individual participants [10] [11]. In addition, the results reported represent an integrated view of the emergent temporal dynamics of collective attention, placing its study within the framework of interdependence and coevolution of human-meme ecosystems. As such, the methodology used here also opens the path to model socio-technical systems through mechanisms that can give raise to mutualistic systems.
[1] F. Wu, B. A. Huberman, *Proceedings of the National Academy of Sciences* **104**, 17599 (2007).

[2] R. Crane, D. Sornette, *Proceedings of the National Academy of Sciences* **105**, 15649 (2008).

[3] J. Leskovec, L. Backstrom, J. Kleinberg, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2009), pp. 497–506.

[4] J. Lehmann, B. Gonçalves, J. J. Ramasco, C. Cattuto, *Proceedings of the 21st international conference on World Wide Web* (ACM, 2012), pp. 251–260.

[5] C. Cattuto, V. Loreto, L. Pietronero, *Proceedings of the National Academy of Sciences* **104**, 1461 (2007).

[6] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, V. Loreto, *Proceedings of the National Academy of Sciences* **106**, 10511 (2009).

[7] B. Gonçalves, N. Perra, A. Vespignani, *PloS one* **6**, e22656 (2011).

[8] L. Weng, A. Flammini, A. Vespignani, F. Menczer, *Scientific Reports* **2** (2012).

[9] J. Bascompte, P. Jordano, C. J. Melián, J. M. Olesen, *Proceedings of the National Academy of Sciences* **100**, 9383 (2003).

[10] J. Bascompte, P. Jordano, J. M. Olesen, *Science* **312**, 431 (2006).

[11] U. Bastolla, *et al.*, *Nature* **458**, 1018 (2009).

[12] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, Y. Moreno, *Scientific Reports* **1**, 197 (2011).

[13] S. Fortunato, *Physics Reports* **486**, 75 (2010).

[14] M. E. J. Newman, *Phys. Rev. E* **69** (2004).

[15] G. Bell, *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 2531 (2003).

[16] P. P. Staniczenko, J. C. Kopp, S. Allesina, *Nature Communications* **4**, 1391 (2013).

[17] R. M. May, S. A. Levin, G. Sugihara, *Nature* **451**, 893 (2008).

[18] S. Saavedra, D. B. Stouffer, B. Uzzi, J. Bascompte, *Nature* **478**, 233 (2011).

[19] J. M. Kamilar, Q. D. Atkinson, *Proceedings of the National Academy of Sciences* **111**, 111 (2014).

[20] S. Borgatti, M. Everett, *Social Networks* **21**, 375 (1999).

[21] J. M. Olesen, J. Bascompte, H. Elberling, P. Jordano, *Ecology* **89**, 1573 (2008).
[22] T. Petanidou, A. S. Kallimanis, J. Tzanopoulos, S. P. Sgardelis, J. D. Pantis, *Ecology Letters* **11**, 564 (2008).
FIG. 1: **Time-dependent Mutualistic Networks** Bipartite representation of the user-hashtag interaction network at the beginning of the observation period (a). An undirected link is shown whenever a user (here represented by an integer number) authors a tweet containing the corresponding hashtag. Panel (b) sketches the sliding-window scheme, which produces matrices that contain the interactions between users and hashtags starting at time $t_0$ and lasting till time $t_0 + \omega$, with $\omega = t_0 - t_0$.
FIG. 2: Nestedness and modularity exhibit anti-correlated patterns. The central panel shows the evolution of modularity and nestedness. The plot corresponds to normalized values, using maximum nestedness ($z_\lambda = 79.64$) and modularity ($Q = 0.57$) for the whole period. Remarkably, both metrics evolve in opposite trends. Modularity reaches its minimal value around the day that the first camps occur, whereas the nestedness achieves a peak value coinciding with the political movement’s central date –that of the largest demonstrations across the country. Top panels represent four snapshots of the data –encoded in bipartite networks–, rows and columns are sorted in decreasing connectivity order (for an optimal visualization of nested patterns, if they exist). Similarly, lower panels represent the exact same matrices, where rows and columns are sorted module-wise (for an optimal visualization of the community structure, if it exists).
FIG. 3: Topical consistence over time despite user turnover. We track the similarity of the “generalist cores” of the network in time with respect to a fixed gold standard (the core of the network when the maximum of the nestedness is observed). Results for different $w$ (12h in the top panel; 3 days in the lower) show that only hashtags build a stable core, guaranteeing the semantic coherence of the topic across time, whereas the core of users suffers a high rate of turnover, indicating that users are frequently pushed to and from the periphery of the network.
FIG. 4: Modular-to-nested transitions occur at different time-scales. Unfiltered, topic-independent Twitter traffic offers the same evidence as the main example (Fig. 1), provided that a suitable time-scale is examined. In the top panel, $Q$ and $Z_\lambda$ scanned at $w = 1$ day fail to show any clear trend. In the zoomed window, a finer resolution of $w = 60$ minutes again reveals the modular-to-nested transition.
Supplementary Information:  
The nested assembly of collective attention in online social systems  
Javier Borge-Holthoefer, Raquel A. Baños, Carlos Gracia-Lázaro, Yamir Moreno  
1Qatar Computing Research Institute, Doha, Qatar  
2Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza,  
50018 Zaragoza, Spain  
3Department of Theoretical Physics, University of Zaragoza, Zaragoza 50009, Spain  
4ISI Foundation, Turin, Italy  
*To whom correspondence should be addressed: jborge@qf.org.qa, yamir.moreno@gmail.com  

I. DATASETS  
The information ecosystem under consideration stems from two disjoint sets, which correspond to hashtags and users from the online platform www.Twitter.com. Following the analogy with interactions in ecological systems, two species are considered: users and hashtags (memes). This implies that we do not consider follower/following links between users, nor connections between hashtags due, for example, to the co-occurrence in the same tweet. We do not consider either message (tweet) contents, except to extract the hashtags in it: our final aim is to keep record of who-said-what in terms of agents using hashtags.  

A. Spanish collection  
The collection from Spain comprises 521,707 tweets containing at least one hashtag, 22,375 unique hashtags and 78,080 unique users. The observation period ranges from the 25th of April at 00:03:26 to the 26th of May at 23:59:55, 2011, and data was collected by selecting all the tweets containing at least one of a preselected set of 70 hashtags related to the 15M movement with the aim of filtering out only the activity related with this topic. Data collection was carried out by the start-up spanish company Cierzo Development Ltd.
B. UK collection

The UK collection comprises 28,928,528 tweets emitted by a set of 842,745 unique users between the 18th of January at 18:41:56, to the 31st of May at 22:41:56, 2013. The set of unique hashtags in this case is 2,196,934. Unlike the case of the spanish dataset, tweets have been filtered by selecting only those that are geolocalized in the UK and Ireland. In this way, this set provides a raw dataset (only limited by geolocalization) of twitter traffic in which the activity is not filtered by topic, hashtags or users.

C. Data as an evolving bipartite graph

As it is clear in the main text, we attempt to account how the user-hashtag ecosystem changes over time. To this end, we build a sequence of snapshots out of the data. These snapshots have an arbitrary window width \( w \), and adjacent snapshots have an overlap of \( \delta w \). Such overlapping scheme is a rather standard procedure when considering chunked temporally-resolved information, to provide a smooth account of change in time.

The question remains how these datasets can be suitably represented. The most natural way to map user-hashtag interactions is through a bipartite graph of relations, which in turn corresponds to a rectangular presence-absence matrix \( B_t = \{b_{uh}\} \), where \( b_{uh} = 1 \) if user \( u \) has posted a message containing \( h \), and 0 otherwise. Noteworthy, this implies that only binary values are considered, i.e. the number of interactions between nodes \( u, h \) is not recorded. Besides, we acknowledge that results in the main text are not affected by the chosen window width (results there correspond to \( w = 12 \) hours and 3 days, with overlaps of \( \Delta = 6, 24 \) hours, respectively). See Section IV for more details.

It is also important to highlight that the \( B_t \) matrices may not contain the same nodes across \( t \): as time advances, users join (disappear) as they start (cease) to show activity; the same applies for hashtags, which might or might not be in the focus of attention of users. This volatile situation is quite normal in time-resolved ecology field studies \[21\]?, \[22\], where the accent is placed on the system’s dynamics –rather than individuals. Moreover, the level of turnover in the sequence of data is very informative, as it characterizes how the system renews its structure over time (see the main text).
D. Pruning the data

The large size of our two datasets –78,081 unique users and 22,376 unique hashtags in the 15M dataset, and 842,745 users plus 4,217,530 hashtags in the UK dataset– handicaps the data processing and makes the calculations time-consuming. We must therefore apply some restrictions to the number of users and hashtags considered in the network.

To this end we apply a rather straightforward criterion, by which we prune the least active users in the data. This means that only top-contributors (and their associated hashtags) show up in the matrices that we study. In doing so, we guarantee that the whole approach makes sense: only by including the most active users we make sure that generalists and specialists will show up –if any nested patterns are to be found. Also the probability of obtaining a connected matrix is higher. Again, we acknowledge that ours is an arbitrary decision. To provide solid evidence, we have tried several matrix sizes.

a. Spain dataset Whereas results reported in the main text are based on the 1024 most active users, we have also tested smaller sets with qualitatively the same results (see Figure 5). In this Figure, we represent the standardized results for both nestedness (left) and modularity (right). Both magnitudes will be described in detail in the following sections. Three dates are also considered at different moments of the 15M movement: three days before the main campings took place –May 12th–, at the onset of the protests; May 15th itself; and May 19th, when the maximum nestedness is achieved and protests are considered to have reached high levels of visibility. Nestedness curves show a tendency to saturate for large values of the the number of users selected. This flattening is achieved at lower values for earlier dates, being far from saturation on May 19th. In view of these results, we can safely conclude that our sampling procedure, i.e., the restriction to the most active users, does not give rise to a misleading claim about the nested structure organized around the movement formation. Furthermore, if we considered the whole system, we would observe greater levels of nestedness at the peak, leading to a more nested structure than that we are actually seeing. The same argument applies to modularity: the curve corresponding to May 19th shows a decreasing tendency, indicating that the modularity of the whole system is actually smaller than that we are measuring from the sampled set.

b. UK dataset For this dataset, the filter is applied in a slightly different way: the cutoff is applied to both users and hashtags, by choosing the 512 more active users and the
512 most-used hashtags. The reason underlying the additional constraint on hashtags and the smaller number of nodes considered, is the large amount of hashtags used in this dataset: 1,024 users can generate from 2,245 to 13,113 hashtags, depending on the observation time window. Some technical details about the observation period, number of users and hashtags, and time-windows width can be found in Table I.

![Graph](image)

**FIG. 5:** Robustness against matrix size. For the $w = 12h$ set some days have been selected. We perform the null model analysis for different cutoffs in the number of users (x axis), and show how the standardised leading eigenvalue (left) and the standardised modularity (right) evolve. The end at $\sim 700$ users for the curve corresponding to $D = 12M$ indicates that the largest possible matrix has been reached, i.e. there are no more active users at that particular day. These results not only guarantee that our conclusions about the nested structure around the 15M are robust, but also show that the observed peak would be more prominent if we considered the real matrix including all the users and hashtags.

As for how we build bipartite networks for the UK dataset, different possibilities arise: on the one hand, we could randomly select a subset of users and hashtags involved in the network, but in this way we might be missing the relevant agents thought to play a major role in the contribution to the nestedness of the whole system. Besides, a random selection could lead to empty matrices (none of the selected users tweeted any of the selected hashtags). We must, nevertheless, remark that this situation is highly unlikely for the 15M event, as a result of the very nature of the dataset: only people and hashtags related to this particular topic were extracted from Twitter. We will be making use of this method as a way to compare the nestedness levels in the 15M with a topic null model, built from data from the
|                | 15M, 2011       | UK, 2013        | UK (inset), 2013 |
|----------------|-----------------|-----------------|------------------|
| Date range     | 25/04 to 26/05  | 18/01 to 31/05  | 31/01 to 06/02   |
| # total users  | 78,081          | 842,745         | 122,553          |
| # total hashtags | 22,376         | 4,217,530       | 264,291          |
| Filter to users | 1,024           | 1,024           | 512              |
| Filter to hashtags | –              | 1,024           | 512              |
| Final # users  | 17,202          | 50,091          | 37,174           |
| Final # hashtags | 12,384         | 19,905          | 22,933           |
| Time windows   | 12h; 72h        | 24h             | 60m; 120m        |
| Overlap        | 6h; 24h         | 12h             | 30m; 60m         |

TABLE I: Datasets summary details. The date range, number of total users and hashtags are displayed. We also indicate the cutoff in the number of users and hashtags (if any) that has been applied. An unspecified hashtag filter indicates that the the hashtag set is determined by the set of selected users. Users are filtered by activity and hashtags by usage. We also show the final number of users and hashtags after the selection process. Finally, the window width and overlap between consecutive windows are also displayed.

UK dataset.

II. NESTEDNESS IN ONLINE SOCIAL NETWORKS

A. Robustness across metrics

Several studies have been focused on quantifying nestedness, the first proposals being made by Hultén [1], Darlington [2] and Daubenmire [3] to describe patterns in which species-poor sites are proper subsets of those ones present at species-rich sites. Nestedness analysis has become very popular among ecologists, and, although the concept is widely accepted, it has not been formally defined, yielding to several distinct metrics [4, 6]. In this work (main text), we adopt a definition numerically confirmed by Staniczenko et al. [13], where nestedness is given by the maximum eigenvalue of the network’s adjacency matrix. This metric is based on a theorem regarding chain graphs first provided by Bell et al. [14], where it is shown that among all the connected bipartite graphs with \( N \) nodes and \( E \) edges, a
perfectly nested graph gives the larger spectral radius. The method is advantageous over other possibilities due to the invariance of eigenvalues under matrix permutations, and the remarkably low computation time required to perform eigenvalue calculations, even for large matrices.

Nevertheless, we have checked the validity of our results against the improved metric NODF, defined by Almeida-Neto et al. [15]. This measure is based on two simple properties: decreasing fill (DF) and paired overlap (PO). Assuming that row (column) \( i \) is located at an upper position in the sorted presence-absence matrix from row (column) \( j \), the decreasing fill condition imposes that a pair of rows (columns) can only contribute to the nestedness if the marginal total –the number of interactions a row (column) has– of row (column) \( i \), is greater or equal to the marginal total of row (column) \( j \). In this case, the paired nestedness, \( N_{ij} \), is equal to the paired overlap, i.e., the number of shared interactions between rows (columns) \( i, j \). The metric can be summarized as:

\[
NODF = \frac{\sum_{ij} N_{ij}}{m(m-1)/2 + n(n-1)/2}
\]

where

\[
N_{ij} = 0 \quad \text{if} \quad MT_i < MT_j
\]
\[
N_{ij} = PO_{ij} \quad \text{if} \quad MT_i \geq MT_j
\]  

Both metrics are compared in Figure 6 (left panel). In the x-axis the standardised value of the leading eigenvector is displayed against the standardised NODF measure. Matrices involved in the plot correspond to graphs at the distinct snapshots with time-window \( w = 1d \). These results are displayed along with the Pearson and Spearman coefficients and their p-values, showing a good linear correlation with p-values \( p < 10^{-3} \).

B. Robustness across significance tests.

The fact that real matrices are usually far from being perfectly nested, imposes the use of a test for the significance of nestedness values. Such a test implies the implementation of a null model and the computation of standardized results, and additionally, allows one to compare matrices with distinct sizes, this comparison being impossible otherwise. Regarding modularity, the metric already includes in its very definition a null model, in such a way
FIG. 6: Comparison against nestedness metrics (left panel) and modularity detection methods (right panel). For every matrix from the set of time windows with $w = 12h$, the standardised leading eigenvalue, $Z_{\lambda}$, and the standardised NODF metric, $Z_{\text{NODF}}$, are computed. There is a good agreement between both metrics, as the Pearson coefficient, $r$, and the Spearman coefficient, $r_s$, show along with their p-values. We also display a comparison of different community detection methods (right). Modularity values from the fast greedy method by Clauset et al.\cite{16} are displayed along the X axis. On the Y axis different methods are presented: $Q_{LE}$ stands for the Newman’s leading eigenvector method\cite{18}, $Q_{EB}$ for the edge betweenness method by Girvan and Newman\cite{19}, $Q_{LP}$ refers to the label propagation method developed by Raghavan et al.\cite{20}, $Q_{IM}$ to the infomap originally developed by Rosvall et al.\cite{21}, and $Q_{ML}$ for the multilevel algorithm by Blondel et al.\cite{20}.

that the modularity obtained is already a comparison with a randomized counterpart of the network.

Different null models may be proposed. For example, one could think of a null model rewiring the set of links present in the network. A strict application of such scheme would not maintain the bipartite structure of the network, and for that reason it should be avoided.

Within this restriction we can still think of some variations. Here we explore two different possibilities, as discussed in \cite{22}. In null model I, the number of links in the network is preserved, but placed at random within the matrix –although respecting the class of the origin and end of it. The degree sequence is therefore not preserved. Null model II is a probabilistic null model where an interaction between hashtag $h$ and user $u$ is established.
with probability proportional to their connectivity,

\[ p_{uh} = p_{hu} = \frac{1}{2} \left( \frac{k_h}{m} + \frac{k_u}{n} \right) \]  

(3)

In the above expression, \( n \) stands for the total number of users, i.e., the first dimension of \( B_{uh} \), and \( m \) for the number of hashtag, equal to the second dimension of \( B_{uh} \). \( k_u \) and \( k_h \) correspond to the degree of user \( u \) and hashtag \( h \), respectively. This model maintains the number of interactions per class only approximately, i.e. it probabilistically maintains the observed total number of interactions.

We can go further and consider an X-swap scheme null model III, in which a rewiring of the edges is applied but keeping constant the degree sequence of the nodes in the system. This null model, however, is too restrictive, and gives a small number of possible configurations, specially for those matrices having few non-empty cells. We must consider null models having a balance between the number of possible configurations and strictness. For this reason we choose to discard null model III, and apply the probabilistic null model (II), which is the strictest between models I and II.

Figure 7 shows a comparison between them. For matrices in the set of time-windows \( w = 12h \), standardised values are displayed. In addition, we also show \( f(x) = x \), to highlight the limit for which the two null models would give identical results. Data from null model II is always well above this limit regarding the nestedness.

III. MODULARITY

Several algorithms, based on the concept of modularity \( Q \), have been proposed over the last decade to detect modular structures in networks. They all aim at identifying the mesoscale organization of networks, that reveals many hidden features invisible from a global perspective of the network, and rely on the detection of densely connected subgraphs. We have tested the most relevant methods for general graphs, and finally applied a fast-greedy algorithm \cite{16} after considering several possibilities. Although some methods designed specifically for bipartite networks have been proposed \cite{13} , they have not been already implemented, to our knowledge, in the common libraries of complex networks, leading us to opt for the conventional algorithms, which are on the other hand, generally accepted and used.
Figure 6 (right panel) shows the modularity values corresponding to matrices in the time-window \( w = 12h \) set. We have checked the fast greedy results (x-axis) against values from different methods (y-axis). Both multi-label and fast-greedy metrics, based on the optimization of modularity, lay on the bisector \( f(x) = x \), and give similar results, yielding the highest modularity values among the rest of the proposals.

IV. TEMPORAL ROBUSTNESS (ACROSS WINDOW WIDTHS)

Beyond assessing the robustness of the results for the nestedness values (regarding the used metrics and null models), we also need to test for robustness against the (admittedly arbitrary) choice of a window-width. This applies both to the soundness of the results in nestedness and modularity.

Here, beyond our reported 12 and 72 hours schemes, we deliver results on nestedness and modularity for additional window sizes, providing evidence that the results in the main text are robust across different time scales: 6 and 24 hours (see Figure 8; \( \Delta t = 12 \) has been left as a reference for the sake of comparison). As is the case of the main text, the overlap between two successive windows is half the size of such window. Upon inspection, it is clear
that results are noisier the narrower the window is—the regularity of the peaks suggests that the measures are sensitive to different circadian rhythms (periodic temporal patterns). For values aggregating the activity for one day and beyond, such periodic variations disappear.

![FIG. 8: Robustness against window width. Standardised nestedness, $Z_N$, and modularity, $Q$, values are displayed for window sizes 6h, 12h and 24h.](image)

V. CORE-PERIPHERY

Meso-scale structures in networks have received considerable attention, as the detection of this intermediate-scale structures can reveal important characteristics that are hidden at both local and global scales. Among the wide diversity of methods aiming at the detection of such structures, community detection methods have become very popular and successful. As it has been said above, they are based on the identification of densely connected subgraphs called communities. However, in this section we focus our attention on a different type of meso-scale structure, known as the core-periphery structure, that helps one to visualize which nodes of the graph belong to a densely connected component or core, and which of them are part of the network’ sparsely connected periphery. Nodes belonging to the core should be relatively well connected to other nodes in the network, regardless of whether they are core nodes or peripheral nodes, whereas nodes in the periphery should be those elements poorly connected with the core, and disconnected from the periphery. According to this intuitive notion, many methods have been proposed. We follow here a method developed by Della Rossa et al.\cite{17}, based on the profile derived by a standard
random walk model. It and can be obtained in a very general framework and is applied here for undirected unweighted networks.

Let \( w_{ij} = w_{ji} \) be the link of weight 1 between nodes \( i \leftrightarrow j \) in our network of size \( N \). At each time step, the probability that the random walker at node \( i \) jumps to node \( j \) is given by \( m_{ij} \):

\[
m_{ij} = \frac{w_{ij}}{\sum_h w_{ih}} = \frac{1}{k_i}
\]

where \( k_i \) is the degree of \( i \). The asymptotic probability of visiting node \( i \) has the closed form

\[
\pi_i = \frac{k_i}{\sum_j k_j}
\]

The method starts by randomly selecting a node \( i \) among those with the weakest connectivities, and assigning \( \alpha_i = 0 \). \( P_1 \), the set of nodes that are already assigned at step \( k \), is then filled with \( i \), \( P_1 = \{i\} \). For the following steps, \( k = 2, 3, \ldots, n \), the node \( j \) attaining the minimum in

\[
\alpha_j = \min_{h \in N \setminus P_{k-1}} \frac{\sum_{i,j \in P_{k-1} \cup \{h\}} \pi_i m_{ij}}{\sum_{i,j \in P_{k-1} \cup \{h\}} \pi_i} + \sum_{i \in P_{k-1}} \left( \pi_i m_{ih} + \pi_h m_{hi} \right)
\]

is selected. If it is not unique, a randomly chosen node among them, \( l \), is selected, and \( P_k = P_{k-1} \cup l \). Although the algorithm presents some randomness, it has been verified that the effect in the analysis of real-world networks is negligible. The core-periphery profile is then the set \( \{\alpha_k\} \), with \( 0 \leq \alpha_k \leq 1 \), where \( \alpha_k = 0 \) for nodes belonging to the periphery and \( \alpha_k > 0 \) for nodes in the core.

As the goal of this section is to identify the possible formation of a core during the days preceding the 15M, a distance metric should be defined. As a first approach we consider the distance between two core-periphery structures as the product between the two \( \{\alpha_k\} \) sequences,

\[
\vec{\alpha_1} \cdot \vec{\alpha_2} = \sum_{i=0}^{n \times m - 1} \alpha_{1,i} \alpha_{2,i}
\]

Notice that the \( \alpha \) “vectors” do not necessarily share the same coordinates, that is, it may happen that a given node (now by nodes we refer to users or hashtags indifferently) present
in $\vec{\alpha}_1$ does not appear in $\vec{\alpha}_2$ because it was not part of the network. Whenever this is the case, we consider the contribution to the dot product to be zero (i.e., as if it were at the periphery). On the other hand, we normalize the above expression in order to get a bounded value: $0 < \vec{\alpha}_1 \cdot \vec{\alpha}_2 < 1$,

$$\vec{\alpha}_1 \cdot \vec{\alpha}_2 = \frac{1}{\|\vec{\alpha}_1\| \|\vec{\alpha}_2\|} \sum_{i=0}^{n \times m - 1} \alpha_{1i} \alpha_{2i}$$  \hspace{1cm} (9)$$

In the main text we have discussed the conformation of a relatively stable core of hashtags around the 15M day, in contrast to a high turnover of users coming to and leaving the core at different snapshots. Here, we scrutinize further such a finding by ruling out the possibility that it could be due, for example, to the fact that the set of users in the core could be similar over the distinct time-windows and change abruptly at the reference point under consideration. To this aim, we additionally measured the distance of a given core from the previous core –the core present in the previous time-stamp. Results in Figure 9 reject this conjecture: the turnover of users is still high –the distance is small– when the core is compared with that in the preceding graph, suggesting that users are actually entering and exiting the key positions in the network. In contrast, hashtags keep relatively constant at high distances, indicating that the core near the 15M is formed smoothly.

VI. ANTI-CORRELATION BETWEEN NESTEDNESS AND MODULARITY

VII. “TOPIC” NULL MODEL

We have discussed in the beginning of this supplementary information distinct possibilities regarding the construction of presence-absence matrices that describe the set of interactions in our systems. We have also mentioned that different cutoffs to hashtags and/or users can be applied, and discussed the more reasonable way to proceed to study nestedness and modularity, which consists of either considering the most active users and their related set of hashtags, or the set of most active users plus most tweeted hashtags at a time. Also, on the side of statistical soundness, we have delved into different null model possibilities.

Now however, our concern focuses on the singularity of the results themselves. In particular, we want to test whether the modularity-nestedness crossover we have observed for particular topics is universal –and in this sense uninteresting– to all the activity on Twitter,
FIG. 9: The figure emulates the results in Figure 3 of the main text; however, distances are computed between each time stamped core and the former configuration. As in the main text, the figure illustrates that hashtags build a more stable core in comparison to users.

or rather it is a specific mechanism underlying the formation of consensus around related information. Thus we explore here three additional possibilities for the $w = 12h$ time-window on the UK dataset. In option (a) we select randomly and independently 512 users and 512 hashtags, and build the corresponding presence-absence matrix. Although the way in which nodes are selected can produce empty matrices corresponding to graphs with no links, this never happened in our dataset (all matrices have more than 20 non-empty cells). In model (b), 512 users are randomly selected and they determine the set of hashtags to consider. Model (c) is analogous but selecting randomly the 512 hashtags to be included, along with the set of users that tweeted them.

These three sets, (a), (b) and (c), can be considered as an additional category of null models that allow us to discern if the nested patterns previously observed are significant: for example, if set (a) showed high levels of $Z_\lambda$ we would not be able to conclude that the
FIG. 10: Anticorrelation between nestedness and modularity. Left panel: 15M dataset. Right panel: UK dataset. For each matrix corresponding to a specific snapshot we display the standardized nestedness (x axis) and the modularity (y axis). Results show a negative correlation between both magnitudes. We measure this correlation through the Pearson coefficient, obtaining the results shown in Table ??.

coordination phase observed in the 15M is relevant, as we would be finding nested patterns even for structures randomly filtered. A comparison between the three methods is displayed in Figure 11. Results include data from Figure 4 (bottom panel) in the main text. We observe that, when we consider independent users and hashtags at random –set (a)–, nested patterns do not show up and the bipartite network do not present any kind of organized structure. The exception is, perhaps, the region between the 3rd of February afternoon and the 4th of February, when the XLVIII Super Bowl took place, probably due to the high relevance of this tournament (if it became global trending topic, even a randomly built network would show, to some extent, a nested structure). When users (hashtags) are randomly selected, but the set of hashtags (users) is closely related to them, the nestedness increase –sets (b) and (c)–, but this is a systematic shift rather than a differential change.

Appendix. Some selected hashtags

In Tables ?? and ?? we display some of the hashtags used in our dataset, along with the number of counts registered.
FIG. 11: Some results for unfiltered Twitter traffic (2013). Set (a) corresponds to our UK dataset with 512 hashtags and 512 users randomly and independently selected. Such a random selection implies that the presence-absence matrix might be empty, although it never happened in this case. In set (b), 512 users have been randomly chosen determining the set of hashtags. Inversely, set (c) have been obtained by randomly filtering 512 hashtags their related users. Finally, set (d) comprises the 512 most active users and the 512 most used hashtags for comparison.

[1] Hultn E: Outline of the history of Arctic and Boreal biota during the Quaternary period, *Thule*: Stockholm, 1937
[2] Darlington PJ: *Zoogeography: the geographical distribution of animals*, Wiley, 1957
[3] Daubenmire R: *Floristic plant geography of estern Washington and northern Idaho*, *J. Biogeogr* 2, 1-18, 1975
[4] Patterson BD, Atmar W: *Nested subsets and the structure of insular mammalian faunas and archipelagos*, *Biol. J. Linn. Soc.* 28, 65-82, 1986
[5] Wright DH, Reeves J: *On the meaning and measurement of nestedness of species assemblages*, *Oecologia* 92, 416-428, 1992
[6] Atmar W, Patterson BD: *The measure of order and disorder in the distribution of*
species in fragmented habitat, *Oecologia* 96, 373-382, 1993

[7] Cutler AH: Nested biotas and biological conservation: metrics, mechanisms, and meaning of nestedness, *Landscape Urban Plann.* 28, 73-82, 1994

[8] Lomolino MV: Investigating causality of nestedness of insular communities: selective immigrations or extinctions?, *J. Biogeogr.* 23, 699-703, 1996

[9] Wright DH et al.: A comparative analysis of nested subset patterns of species composition, *Oecologia* 113, 1-20, 1998

[10] Brualdi RA, Sanderson JG: Nested species subsets, gaps, and discrepancy, *Oecologia* 119, 256-264, 1999

[11] Cam E et al.: Inferences about nested subsets structure when not all species are detected, *Oikos* 91, 428-434, 2000

[12] Hausdorf B, Henning C: Nestedness of north-west European land snail ranges as a consequence of differential immigration from Pleistocene glacial refuges, *Oecologia* 135, 102-109, 2003

[13] Staniczenko PA, Kopp JC, Allesina S: The ghost of nestedness in ecological networks, *Nature Communications* 4 1391, 2013

[14] Bell FK, Cvetkovi D, Rowlinson P, Simi SK: Graphs for which the least eigenvalue is minimal, II, *Linear Algebra Appl* 429, 2168-2179, 2008

[15] Almeida-Neto M, Guimarães P, Guimarães Jr PR, Loyola D, Ulrich W: A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement, *Oikos* 117 1227-1239, 2008

[16] Clauset A, Newman MEJ, Moore C: Finding community structure in very large networks, *Phys Rev E* 70, 066111, 2004.

[17] Della Rossa F, Dercole F, Piccardi C: Profiling core-periphery network structure by random walkers, *Scientific Reports*, 3, 1467, 2013

[18] Newman MEJ: Finding community structure in networks using the eigenvectors of matrices, *Phys Rev E* 74 036104, 2006

[19] Girvan M, Newman MEJ: Community structure in social and biological networks, *Proc. Nat. Acad. Sci. USA* 99, 7821-7826, 2002

[20] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E: Fast unfolding of community hierarchies in large networks, *J Stat Mech* P10008, 2008
TABLE II: Modularity values obtained for different algorithms implemented in the igraph library: LE: leading eigenvector, RW: random walk (walktrap), EB: edge betweenness, FG: fast greedy, ML: multilevel, LP: label propagation, IM: infomap. FG gives in all cases the best modularity value.

| n  | m | Q_{LE} | Q_{RW} | Q_{EB} | Q_{FG} | Q_{ML} | Q_{LP} | Q_{IM} |
|----|---|--------|--------|--------|--------|--------|--------|--------|
| 90 | 4 | 0.4571 | 0.4363 | 0.4571 | 0.4608 | 0.4608 | 0.4325 | 0.4603 |
| 238| 8 | 0.5705 | 0.5635 | 0.5697 | 0.5726 | 0.5650 | 0.5393 | 0.5646 |
| 534| 16| 0.5175 | 0.5104 | 0.5447 | 0.5492 | 0.5365 | 0.5375 | 0.5432 |
| 939| 32| 0.4478 | 0.4720 | 0.4926 | 0.5067 | 0.4905 | 0.4236 | 0.4874 |

TABLE III: Pearson correlation coefficient values for sets displayed in Figure 10, along with their p-values. Results show a high correlation above $-0.94$ in the case of the 15M, and a slightly smaller value for the UK dataset. However, if we remove the set of points located at $10 < Z_\lambda < 15$ and $Q \sim 0.90$, the Pearson correlation increases up to $-0.92$.

| Dataset | $r$  | p-value |
|---------|------|---------|
| 15M, $w = 12h$ | -0.94 | 0.00 |
| 15M, $w = 3d$  | -0.95 | 0.00 |
| UK, $w = 1h$   | -0.83 | 0.00 |
| Hashtag                  | Counts  | Hashtag                  | Counts  |
|-------------------------|---------|-------------------------|---------|
| #acampadasol            | 189251  | #acampadasol            | 5760    |
| #spanishrevolution      | 158487  | #acampadasol            | 5760    |
| #nolesvotes             | 66329   | #22m                    | 5205    |
| #15m                    | 65962   | #reflexion              | 4693    |
| #nonosvamos             | 55245   | #sinbanderas            | 4481    |
| #democraciarealya       | 47463   | #consensodeminimos      | 4348    |
| #notenemosmiedo         | 32586   | #italianrevolution      | 3981    |
| #yeswecamp              | 31811   | #estonosepara           | 3860    |
| #acampadabcn            | 20069   | #acampadaalicante       | 3593    |
| #15mani                 | 17986   | #tomalacalle            | 3517    |
| #acampadasevill1a       | 14356   | #fb                     | 3372    |
| #globalcamp             | 13186   | #europeanrevolution     | 3035    |
| #acampadavalencia       | 13129   | #acampadapamplona       | 2839    |
| #estoesreflexion        | 11080   | #worldrevolution        | 2777    |
| #acampadagranada        | 9717    | #democraciarealya       | 2766    |
| #acampadamalaga         | 6808    | #acampadapalma          | 2709    |

TABLE IV: Top 32 most-used hashtags in the 15M dataset.
| Hashtag  | Counts  | Hashtag  | Counts  |
|----------|---------|----------|---------|
| #london  | 127082  | #essex  | 38043   |
| #ff      | 110972  | #cbb    | 37741   |
| #uk      | 80238   | #lol    | 37646   |
| #love    | 70618   | #tired  | 37316   |
| #jobs    | 70268   | #legend | 37225   |
| #weather | 66366   | #summer | 36739   |
| #excited | 64795   | #cute   | 33648   |
| #mufc    | 60176   | #bgt    | 33586   |
| #endomondo | 55537  | #coys   | 33380   |
| #xfactor | 51870   | #wimbledon | 32510  |
| #lfc     | 47237   | #nufc   | 32328   |
| #stalbans| 45117   | #help   | 32118   |
| #mtvhottest | 43572  | #amazing | 30124  |
| #happy   | 40086   | #bored  | 30066   |
| #loveit  | 40049   | #bbuk   | 29960   |
| #nowplaying | 39798  | #ukweather | 29940 |

TABLE V: Top 32 most-used hashtags in the UK dataset.